NPS-OR-11-003 Rev.

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

HUMAN FACTORS IN THE JOINT TYPHOON WARNING CENTER WATCH FLOOR

by

Eva Regnier and Alex Kirlik

March 2011: Revised November 2012

Approved for public release; distribution is unlimited

Prepared for: Joint Typhoon Warning Center
425 Luapele Road
Pearl Harbor, HI 96860

THIS PAGE INTENTIONALLY LEFT BLANK

NAVAL POSTGRADUATE SCHOOL

Monterey, California 93943-5000

The report entitled *"Human Factors in the Joint Typhoon Warning Center Watch Floor"* was prepared for and funded by Joint Typhoon Warning Center, 425 Luapele Road, Pearl Harbor, HI 96860.

Further distribution of all or part of this report is authorized.

This report was prepared by:

Eva Regnier Visiting Associate Professor Operations Research	Alex Kirlik Consultant

Reviewed by:

Ronald D. Fricker Associate Chairman for Research Department of Operations Research	Robert F. Dell Chairman Department of Operations Research

Released by:

Jeffrey D. Paduan
Vice President and Dean of Research

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved* OMB No. 0704-0188

1. REPORT DATE ***(DD-MM-YYYY)*** 30-11-2012	**2. REPORT TYPE** Technical Report	**3. DATES COVERED** ***(From-To)*** 01-10-2010 – 30-03-2011

4. TITLE AND SUBTITLE Human Factors in the Joint Typhoon Warning Center Watch Floor	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Eva Regnier and Alex Kirlik	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943	**8. PERFORMING ORGANIZATION REPORT NUMBER** NPS-OR-11-003Rev.
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Joint Typhoon Warning Center 425 Luapele Road Pearl Harbor, HI 96860	**10. SPONSOR/MONITOR'S ACRONYM(S)** NMFC/JTWC
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES
The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

14. ABSTRACT

This study evaluates the task and support environments associated with the Joint Typhoon Warning Center (JTWC) watch floor and provides recommendations to improve forecast accuracy. The principal findings indicate that, at this time, factors in the task environment are very likely limiting forecast accuracy. In particular, system-induced practical and cognitive limits on the forecaster's repeatability, i.e., the ability to reproduce an identical forecast given identical information, limits forecast accuracy. Thus, forecasters' performance and forecast accuracy could be enhanced by an information integration system with recommended features, by more precise standard operating procedures, and by training and feedback better matched to the task. Further studies are also recommended.

This version rescinds and replaces prior versions. The appendix has been corrected.

15. SUBJECT TERMS
human factors, forecasting, tropical weather

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Eva Regnier
Unclassified	Unclassified	Unclassified	UU	57	**19b. TELEPHONE NUMBER** ***(include area code)*** 831-656-3461

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The objective of this project is to evaluate the task and support environments associated with the Joint Typhoon Warning Center (JTWC) watch floor and provide recommendations, with the ultimate goal to improve forecast accuracy.

In support of this objective, Dr. Alex Kirlik (AK) and I visited the JTWC during the West Pacific typhoon season in November 2010, to observe, conduct interviews, and review documents related to the task environment of the forecasters and satellite analysts. AK was present at the JTWC on November 12 and November 15-19, while I was present November 15-18.

AK's report is attached as an appendix. He finds that, at this time, factors in the task environment are very likely limiting forecast accuracy. In particular, system-induced practical and cognitive limits on the forecaster's repeatability, i.e., the ability to reproduce an identical forecast given identical information, limits forecast accuracy. Thus, forecasters' performance and forecast accuracy could be enhanced by an information integration system with recommended features, by more precise standard operating procedures, and by training and feedback better matched to the task. Developing alternative performance metrics would be relatively low-cost and can be implemented fairly quickly. If used internally and externally, these metrics have the potential to improve both forecast accuracy and user satisfaction.

The remainder of this section gives findings and recommendations, with references to AK's report. Section II discusses performance evaluation metrics, why they are an important part of AK's findings, and provides a supplementary explanation of the effect of inconsistency on mean absolute track error. Section III highlights the problems we identified with current assessments of the human contribution to the official forecast, as well as the potential value of a more accurate quantification of this contribution, and possible approaches to its measurement. While AK's report focuses on the track forecast, Section IV describes special challenges to forecasting intensity.

A. PRINCIPAL FINDINGS

1. The human forecaster[1] is squeezed by the time and mental workload requirements of the data collection (up to five steps to load a single image) and integration[2] process at the beginning of the forecasting cycle and forecast production at the end of the forecasting cycle. In the middle, there is a restricted time for the thinking portion of the process (including comparing model output with observations, evaluating model behavior, and assessing the current state of the atmosphere and its likely future

[1] Hereafter, any member of the watch team, including Typhoon Duty Officer (TDO), Typhoon Duty Assistant (TDA), and satellite analyst will be referred to as forecasters.

[2] This is meant to capture synthesizing information mentally, e.g., mental overlays, awareness of relative importance/quality, but not the step of processing it or integrating it to produce judgments/assessments.

evolution), where we expect the human-in-the-loop (HITL) can add the most value. Moreover, there are often distractions and competing demands on the forecasters' time throughout the cycle, such as calls from customers. This indicates that forecasters would perform better if the time requirement and mental workload were reduced in the data-collection and integration phase and the forecast-production phase. Recommendations 1, 2, 6, and 7 in Section I.B respond to this finding.

2. As expected, the forecasters' task environment is information rich. The quantity of information available and potentially available may itself be limiting forecasters' performance through information overload. Perhaps more important, there are substantial practical barriers to forecasters' acquiring and synthesizing the relevant information (guidance) and meta-information (e.g., age, and validity of images and models) that unnecessarily restrict the time available for forecasters to assess the current state of the system. In addition, it's known that human experts (presumably including JTWC) are more accurate when the information acquisition is separated from its analysis (see citations in Stewart & Lusk, 1994). Recommendations 1, 2, 3, 6 and 8 in Section I.B respond to this finding.

3. The primary verification metrics used by the JTWC may be an impediment to improving forecast accuracy. The mean absolute track error and mean absolute intensity error metrics match neither:

   (a) forecasters' mental process, in which they assess future speed and direction of a storm, nor

   (b) customers' decision criteria, e.g., tropical cyclone (TC) conditions of readiness (TCCOR) definitions or other considerations affecting user acceptability.

   Recommendations 3, 4, 5, and 9 in Section I.B respond to this finding.

4. Currently, there is no way to evaluate the contribution to forecasts provided by HITLs. Often the mean track error for the consensus track (CONW) is compared with the mean track error for the JTWC official (OFCL) track. However, the CONW benefits from HITL intervention at two or more points in its generation, and therefore does not represent a purely automated product. Recommendations 6 and 7 in Section I.B respond to this finding.

B. RECOMMENDATIONS

1. **Improved information integration system.** As AK recommends (Section 5-3-1, p. 32), a high-value intervention would be acquisition of an information integration system and interface that makes information acquisition faster and with few, if any, actions required by the forecaster. Section 2.3 of AK's report includes details on interface features that research shows would support enhanced forecaster performance, including consistent and automatic geo-temporal overlays such that guidance from multiple information sources are matched geographically and temporally. An additional feature that might help improve forecaster accuracy would be visual indications of the validity (predictive contribution) of guidance when that validity changes from one forecast scenario to another. For example, for the satellite analyst, the age of various images and automated products may vary across forecast cycles; the more recent each product is, the higher its validity. The same applies to numerical model tracks.

2. **Detailed standard operating procedures for forecast process.** A further intervention that would tend to reduce human-induced inconsistency would be detailed standard operating procedures for the forecast process, perhaps in the form of checklists (see Appendix, Section 5-3-2, informed by results of additional studies as described in Appendix 5-5 and Recommendations 6 and 9).

3. **Training, feedback and interface features matched to the forecasting task.** Matching all elements of the forecast environment to the mental process, with consideration of the relationship between the speed and direction errors and track error (FTE), could improve forecast accuracy (see Appendix, Sections 3, 5-1, and 5-2).

   The timing of recurvature is another forecaster assessment that, with speed and direction, determines track positions. This should be considered as an element of training and the design of diagnostic and feedback metrics (see Recommendation 4).

   In addition to synthesizing a large volume of current guidance, forecasters must also subjectively combine the most recent guidance with prior guidance to produce a forecast. AK recommends (Appendix, Section 5-6) that training and perhaps additional relevant meta-information could improve their performance in this difficult task.[3]

4. **Alternative performance metrics.** The mismatch between the primary verification metrics and the forecasting task as well as customers' needs suggests that the JTWC may want to adopt alternative performance metrics that provide more relevant and therefore effective feedback to the forecasters, and potentially to customers. These metrics may be used primarily for training, feedback, and diagnosis, in particular if existing metrics are mandated external to the organization. If well-designed and used, they could still improve the accuracy of the existing mean absolute error metrics (see Appendix, Sections 3 and 5-7, and Section II below).

5. **Consideration of alternative product formats.** AK recommends alternative product formats to improve communication of uncertainty in the track forecast (Appendix, Section 5-7). In addition to the issues discussed by AK, alternative product formats could address the fundamental mismatch between users' information requirements and interpretation process and the JTWC's forecast fields. The language of TCCOR definitions provides one (of many) examples of this mismatch.

The following are recommendations for further studies.

6. **Study the effect of a go-with-CON feature.** AK recommends (Sections 5-4 and 2.4, Item 5) implementing a feature in the automated tropical cyclone forecasting system (ATCF) that allows the forecaster to choose to allow the system to automatically generate an OFCL forecast based on the automated guidance, i.e., the CONW track, the statistical typhoon intensity prediction scheme (STIPS) intensity, and the DRCL CLIPER (climatology and persistence, Knaff et al., 2007) wind radii. This could improve error statistics by eliminating variability due to human factors in situations in which the human forecaster does not expect to be able to add value over and above the automated products, and by reducing the forecaster workload required to "lay-down" the OFCL forecast when the forecaster determines he cannot add value over and above the automated products. Whether the availability of this feature adds value, and under what circumstances deviations add value, should be studied (see Appendix, Sections 5-4 and 5-5).

[3] A related, but distinct consideration is regression to the mean or to climatology. Numerical models' predictive validity falls off quickly with lead time, as does the predictive power of satellite imagery and recent observations. Therefore, the forecast should reflect a combination of guidance and long-run typical storm behavior (climatology), where the weight applied to climatology increases at longer leads. Forecasters' long-lead likely reflect a tendency to long-run typical behavior of storms in the region, which may explain, in part, why the performance of OFCL track relative to CONW increases at longer leads. The models do not regress to climatology, while the human does. However, it is common for human experts to "anchor" (and therefore assign too much weight to) the first guidance they receive, and therefore training to create optimal regression to long-term patterns might improve forecast accuracy, especially at long leads.

The go-with-CON feature could be made the default starting point for the warning product within the ATCF, or it could be a feature that needs to be selected. The ability to deviate from the automated product would nevertheless be preserved. When the forecaster determines that it is appropriate, he would be able to modify any or all forecast fields (positions, intensities, and radii at all leads), and the remaining automated fields would automatically become the OFCL forecast fields.

Alternatively, the automatic forecast could adjust the output from the automated products to improve performance with respect to the user-acceptability metrics discussed in Recommendation 9 and Section II.C. For example, it could reflect a weighted average of the prior official forecast and the automated products. The goal is to automatically produce a forecast that reflects what the forecaster intends to do by combining automated products with persistence, with respect to prior official forecasts, while reducing workload and unintentional (human-induced) variability.

7. **Quantify the effect of the HITL.** As discussed in Finding 4 and Section III, there is currently no way to measure the value of the HITL. In order to understand both how and when the human forecaster adds value, and seek to increase that value, a preliminary step is to measure it. For example, research has shown that in other environments, human experts can add value relative to models in situations of high uncertainty. Therefore, a near-term research question is whether the JTWC's accuracy statistics outperform consensus in certain identifiable situations, such as high uncertainty, or situations reflecting particular patterns of disagreement among the numerical models. Once the value of the HITL can be measured, the next step is to identify situations in which the HITL improves forecast accuracy, relative to automated products.

8. **Conduct Lens Model Analysis of Intensity Forecasts.** Despite improvements in guidance, TC intensity forecast error has not decreased significantly in the last 20 years (DeMaria, Knaff and Sampson, 2007). The task environment faced by TC forecasters has many characteristics known to challenge human judgment. As discussed in Section IV, the intensity forecast may be even more challenging than the track forecast. In the JTWC, these factors are compounded by personnel turnover. A lens model analysis (described in the Appendix, Section 4 and Stewart, 1990, and applied by Stewart, Roebber & Bosart, 1997) of the forecasting and verification processes can quantify these effects and diagnose specific sources of error. This framework has been used to study subjective judgment in many contexts, including aviation and military applications (Kirlik, 2006). In the context of TC intensity forecasting, some possible sources of error are inconsistency in interpreting images and discounting aging guidance, redundancy in information content and cognitive errors

like overweighting assessments of correlated experts or sources. The results may suggest interventions that could improve TC intensity forecast accuracy (Stewart & Lusk, 1994).

9. **Develop and implement user-acceptability metrics.** JTWC's customers care about forecast characteristics other than accuracy. Therefore, it is appropriate to develop measures of user acceptability to capture the value the JTWC adds with respect to these other considerations. User-acceptability metrics, complementary to accuracy metrics, can be used internally to determine whether the OFCL forecast's deviations from automated products are the result of appropriate adjustments and externally to document user value. Examples of forecast characteristics for which metrics could be designed include jerkiness (nonsmoothness) of a given forecast track, jumpiness of a track from update to update, and frequency of changes to TCCOR settings that would be induced by use of the forecast.

II. PERFORMANCE EVALUATION METRICS

AK's report has an extensive discussion of the implications of the primary performance metric used by the JTWC—mean absolute error (in track and intensity). At the beginning of this study, we did not anticipate that some of the most important findings would relate to the verification process. However, performance measurement can have a significant impact on performance. First, forecasters (like other experts) learn from feedback. When the feedback is highly relevant to the task and forecasters can match their measured performance with their experience of their task, their ability to learn from prior performance and improve their performance is enhanced.

Second, metrics create an incentive to perform well with respect to the metric (in other words, "what gets measured gets managed"). Incentives are related to feedback, but distinct. Even if forecasters' primary objective is not fully aligned with the performance metric, they feel rewarded by performing well with respect to the published metrics, which could cause them to slightly and unconsciously adjust their forecasts to improve performance metrics to the detriment of their objective. For example, the forecasters' responsibility to their customers (and therefore their underlying objective) may require them to err on the side of caution or consistency with prior forecasts when assets are threatened, or to produce a meteorologically plausible track. To the extent that they experience a competing incentive to reduce track error (FTE), they may reduce the value of the product to the customer.

A. IMPACT OF INCONSISTENCY

In Section 3.2, AK describes how imprecision (which can be created by human inconsistency) increases mean absolute error. To expand on the possibly counterintuitive point that imprecision adds to error, and does not cancel itself out, an additional explanation follows.

Figure 1 represents a single forecast, at a single lead, and its verification. The best track (verifying) position is shown in black. At the time he issues the forecast, the forecaster does not know the best-track position, but has a best guess, whose location is shown in green. If the forecaster were able to precisely forecast the best-guess position, the track error would be R, the distance between the best guess and the best track. However, he can't precisely issue the best-guess forecast. The reasons include:

- Interface barriers, e.g., pixilation in the forecasting interface and imperfect hand control of the mouse in laying in a forecast track line;[4] and

[4] In ATCF, the representation of the model and CONW tracks is forecast positions connected by straight lines. However, the representation forecasters (and, eventually, users) get of the OFCL forecast is forecast positions connected by curved lines. When forecasters (or, presumably, users) see the curved lines, they think that's a representation of the track the storm is forecast to follow. For this reason, forecasters sometimes adjust track positions to make those curved lines look more realistic (sometimes they might show a stair-step pattern or other non-meteorological artifact of the way the curves are generated), so the forecaster will massage the positions to get lines that look more realistic.

- Imperfect forecasting repeatability, due to impediments to perfect, machine-like reliability, e.g., fatigue, time pressure, overcaution due to recent forecast bust or potential threat to an asset, day-to-day variations in set of guidance used, order in which they were viewed, lack of time to thoroughly evaluate all guidance.

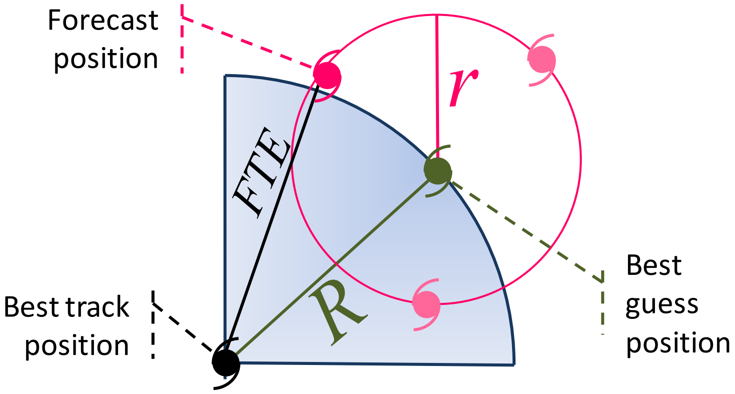

Figure 1: Model of the effect of imprecision on forecast error.

Therefore, the actual forecast issued will not coincide precisely with the best possible guess. In Figure 1, this imprecision is modeled such that the actual forecast issued is a distance r from the best guess position, and any point at distance r (the red circle) is equally likely.[5] Examples of possible forecast positions are shown in red.

While the intuition might be that half the time imprecision will add to the error and half the time it will reduce the error, Figure 1 shows otherwise. All points on the red circle that are outside the blue circle have greater FTE than the best-guess position, and more than half the red circle is outside the blue circle. The forecast position shown in deep pink is an example of a forecast position that's on the half of the red circle closest to the best-track position, yet has larger FTE than the best track position. Since all points on the red circle are equally likely, the probability that imprecision increases FTE is greater than 50%. On average, imprecision increases track error.

B. MISMATCH BETWEEN METRICS AND FORECASTING PROCESS

In Section 3.2, AK describes the mental process that forecasters appear to follow in determining their track forecast—assessing the anticipated speed and direction of the storm, and then turning this into track positions. AK describes the nonlinear and asymmetric (for positive and negative errors) effect of errors in the directly-assessed variables (speed and direction) on FTE.

[5] The general result is unchanged if the distance r is modeled with a probability distribution rather than as a constant, because the effect that imprecision adds to error, on average, holds for every nonzero value of r . If, for some reason, imprecision is distributed such that not all directions are equally likely, then the general result could change.

These insights explain an observed slow bias in OFCL track forecasts. AK's analysis (see, in particular, Figure 3-8) indicates that negative errors in speed will produce smaller FTE than positive errors. This would incentivize forecasters to bias their forecasts on the slow side. Forecasters may have intuitively, or perhaps even consciously, detected this effect, and may be biasing their track and speed assessments accordingly. This supports the belief that verification metrics affect forecaster behavior and interventions that provide incentives and feedback based on the JTWC's most important objectives could improve forecaster performance, with respect to those objectives, discussed in Section II.C.

While AK's figures show FTE as a function of proportional error in speed, the effect holds for absolute distance error as well. In other words, when direction error is less than 90°, predicting a position 10 nautical miles (nmi) too close to the current position produces a lower FTE than predicting a position 10 nmi too far away, and even a lower FTE than a forecast position that is the correct distance (and speed), but the wrong direction.

C. USER ACCEPTABILITY

Although the JTWC summarizes the performance of its official forecast with mean absolute track error and mean absolute intensity error, the forecasting staff is aware that small changes in the official forecast that may reflect insignificant differences in the understanding of the meteorology can have substantial operational impacts. Numerical models and their resulting consensus, on the other hand, are tuned to optimize long-run accuracy statistics described above without consideration of the effects of minor changes on customers' operations. The CONW track is a simple average of a set of meteorologically plausible tracks; it is not constrained to be meteorologically plausible.

However, many of the direct consumers of JTWC products are themselves meteorologists, supporting Navy ships and installations. Their confidence in the product could be reduced by forecasts that don't make meteorological sense. JTWC forecasters produce forecasts that, within the margin of uncertainty with respect to the best meteorological evidence and interpretation, will not tend to undermine user confidence in the product or cause unnecessary consequences to the customer. For example, the JTWC track tends to be smoother than CONW, and is therefore more representative of actual TC paths.

The above enhance user acceptability, in part by encouraging customer confidence in the forecast and forecasters. Another factor that can affect user confidence is large changes between forecast updates. In addition, large changes in the JTWC forecast from warning to warning cause concern and in many cases unnecessary cost among the customers. Therefore, the JTWC will tend to persist with existing forecast until meteorological evidence drives a change.

CONW exploits the best available scientific knowledge from many models, some of which are explicitly tuned to provide highly accurate TC predictions, while being free of any considerations related to user acceptability. Therefore, it is hard to beat CONW in terms of long-run accuracy, and deviations from CONW will tend to degrade the forecast's error statistics.

Since JTWC is subject to considerations other than accuracy, it is appropriate to develop measures of user acceptability to capture the value the JTWC adds with respect to these other considerations. Performance metrics could be designed to quantify the user-acceptability of the OFCL forecast. For example,

- jerkiness (nonsmoothness) of a given forecast track;
- jumpiness of a track from update to update; and
- frequency of changes to TCCOR settings that would be induced by use of the forecast.

The last could be measured by comparing an automated TCCOR-setting rule. For example, set TCCOR 4 (3,2,1) if the forecast track with 50-knot (kt) wind radii would indicate 50-kt winds at a given base within the next 72 (48,24,12) hours. If the following forecast update changes so that the condition is no longer satisfied, that would be counted as a reversal of the TCCOR setting. If the next level is set before the given level is reversed, then this counts as a reversal of a TCCOR. For TCCOR 1, if the storm dissipated without the base actually experiencing 50-kt winds, that would count as a reversal. These results could be summarized as in Table 1.

Table 1: Frequency of TCCOR reversals

TCCOR Level	CONW (+DRCL radii)	JTWC
4	Count these occurrences within a given period for comparison.	
3		
2		
1		

The same data could be used to conduct a signal-detection theory study. Brooks (2004) is an example of signal-detection theory applied to meteorological forecasts; in that case, for tornados.

III. CONTRIBUTION OF HUMAN-IN-THE-LOOP (HITL)

To measure the value provided by the HITL fairly, a forecasting system that includes human intervention should be compared with a system that is fully automated. Current comparisons of error statistics for CONW versus the JTWC official track do not accurately measure the value of all contributions of human intervention in the TC forecasting process.

The human contributes to the CONW during (at least) two stages in each forecast cycle. The human TDO produces a bogus that the numerical models take as an input. The JTWC's position, produced by the TDO with input from the human satellite analyst, is an input to the track interpolation scheme: the tracks generated by the various models are translated so that their current position coincides with the analysis. Some numerical models even use recent JTWC best tracks as inputs. The intensity analysis is also an input to statistical intensity prediction models, including STIPS.

In order to properly measure the value added by the HITL, the official forecast should be compared with a forecast generated by a system that has no human contribution. To estimate this value, you would need to be able to run the models with an automated bogus, and automate the translation of the tracks to the analysis position.

A nearly-automated consensus track (which would still benefit from any human value-added in, producing the bogus) could be either

- an average of noninterpolated models, or
- an average interpolated to an automated best track.

While we would expect the second type of consensus to be more accurate, it is worth comparing both of the above with the JTWC forecast, to at least allow for the possibility that interpolating to an automated best-track degrades forecast accuracy. Thus, more fair measures of the effect of human intervention on track and forecast accuracy statistics would be:

- Track: mean absolute 48-hour position error for automated consensus (type 1 or 2 above) – mean absolute 48-hour JTWC position error.
- Intensity: mean absolute 48-hour STIPS intensity error – mean absolute 48-hour JTWC intensity error.

In recent years (with the exception of 2010 and long-lead forecasts), the CONW track forecast has had lower seasonal average FTE than the OFCL forecast, and STIPS intensity predictions have had lower seasonal average error than the OFCL forecast. These are the most commonly cited measures of forecast performance.

Averaged over an entire season, the accuracy statistics (mean absolute track and intensity error) for the consensus are somewhat better than for the JTWC track (except at long leads). However, it may be that even based on these accuracy statistics, JTWC outperforms the consensus systematically under certain circumstances. For example, research has shown that in other environments, human experts can add value relative to models in situations of high uncertainty. Therefore, a near-term research question is whether the JTWC's accuracy statistics outperform consensus in certain identifiable situations, such as high uncertainty, or situations reflecting particular patterns of disagreement among the numerical models.

For some storms, such as 15W in 2010, the forecaster may have knowledge indicating that certain model tracks are likely to be in error. With respect to intensity, the forecasters are aware that the statistical models, like STIPS, perform well over many storms, but underforecast intensity when there are signs of rapid intensification, as occurred for 15W. In these situations, the human can add value, even as measured by accuracy statistics, by deviating from the automated guidance. The trick is identifying these situations in advance.

A first step is to explore the historical record, and categorize storms according to potential indicators, such as model spread (although prior research has not shown that spread alone does not necessarily indicate an opportunity for the human to intervene). For example, do bifurcated model tracks versus a "squashed spider" indicate a situation in which the TDO should deviate from CONW? More sophisticated indicators, which perhaps can distinguish a bifurcation situation in which two alternative scenarios are reflected in the models, might be better at distinguishing the situations in which humans can improve the forecast. In a bifurcation situation, if the forecaster has access to recent data or knowledge that is not incorporated into the models, he may be able to eliminate those that are not predictive and thereby improve on CONW.

Given the JTWC's experience with the systematic approach forecast aid (SAFA), which automatically flagged certain model behaviors that would be susceptible to change as models changed, the search for patterns of high-HITL value would have to use indicators that are predictors over a long record of storms, covering model changes. In identifying circumstances in which the human adds value, it would also be important to separate out the two or more elements of human contributions, i.e., the satellite analysis and the rest of the forecast.

IV. SPECIAL CHALLENGES FOR INTENSITY

AK's report focused on the track forecast. Most of his observations and recommendations apply equally to the intensity forecasting problem. For example, the impediments to collecting and synthesizing the relevant guidance are at least as great for intensity as for track. Moreover, track and intensity (structure) are both features of a single complex system, compounding the difficulty of the forecasting task environment. Additional challenges not discussed elsewhere that may pose a special challenge for intensity prediction include:

- Feedback is noisy. By definition of intensity, the ground truth—maximum sustained surface wind speed—is rarely, if ever, observed. From the National Oceanic and Atmospheric Administration (NOAA) Science Advisory Board (SAB) majority report from the Hurricane Intensity Research Working Group (2006, p. 10):

  > The National Hurricane Center definition of 'intensity' is the maximum 1-minute-sustained 10-m-height winds in the core of the storm. It provides an easily grasped measure of storm strength. However, this quantity is rarely, if ever, directly measured, and is normally inferred by extrapolation from ground or aircraft observations, by satellite pattern-recognition techniques, or by pressure-deficit/maximum-wind relationships.

  By definition, forecasters are aiming at a hazy target, and because the verification is usually subjective, an additional layer of error (bias, unreliability, and imperfect aid exploitation) is introduced. A further layer of imprecision is introduced by measurement error, especially as the types of measures that may be available to inform the best track analysis of intensity varies from forecast to forecast. In addition, the best track is rounded to 5kts, further adding noise to the best track intensity.

- Environmental unpredictability (low match between true descriptors and the actual event) reduces forecast accuracy directly, but also tends to reduce the forecasters" reliability (Stewart, 2001), to the further detriment of the forecast. Intensity is highly unpredictable; therefore, it is an especially difficult forecasting challenge.

- Feedback is delayed and not visually presented in ATCF. Accurate feedback on track is more immediate and more salient than on intensity, because the analysis center available within about eight hours is very close to the eventual (postseason) best track, and is represented graphically in ATCF, and may be displayed simultaneously with the recent forecast so that differences are apparent and may be viewed in context of the guidance used to produce the forecast.

- Many of the intensity guidance products (both automated and human-mediated) rely on the same underlying images, and therefore have high correlation. Human forecasters tend to attribute too much of the predictive value to multiple correlated cues that actually reflect the same information.

- Many of the guidance products are themselves human-mediated, and therefore subject to imprecision introduced by human factors.

These issues raise some important questions that could be addressed with a lens-model analysis, such as:

- Do the above-described challenges limit intensity forecast accuracy, and if so, by how much?
- How much could perfect guidance (valid forecast aids that, when combined objectively, explain 100% of the variability in verified intensity) improve forecast accuracy in the current TC forecasting environment?
- Does subjective verification limit the maximum potential accuracy of intensity forecasts, and if so, by how much? Could accuracy be improved by using objective verification?

APPENDIX. KIRLIK REPORT

Human Factors at the Joint Typhoon Warning Center (JTWC) Watch Floor

Final Report

Submitted to:

Naval Postgraduate School

Submitted by

Alex Kirlik, PhD
(Independent Consultant/Contractor)
1201 Waverly Drive
Champaign, IL 61821 USA

November 16, 2012

Contents

Executive Summary

The scope of this project was to evaluate the task and technological support environments at the Joint Typhoon Warning Center (JTWC) watch floor and provide recommendations with the ultimate goal to improve forecast accuracy.

Based on a review of available documents, information gathered at the JTWC through observations, interviews, discussions, surveys of information technology and interfaces on the watch floor, a variety of mathematical analyses and related research findings presented in this report, the best available theory of human judgment (forecasting) under uncertainty and optimal belief updating, and best practices in human factors, the following conclusions and recommendations are provided:

1. JTWC staff face severe cognitive challenges associated with attempting to maintain situation awareness of atmospheric state and dynamics from numerous information displays with disparate formats, geo-referencing, measurement scales, symbology, data age, source pedigree, and trustworthiness. Every effort should be made to provide updated technology in an integrated suite of information displays with maximum design consistency.

2. JTWC staff benefit strongly by numerical model guidance, but the cognitive processes whereby they integrate the information provided by guidance and information from their own cognitive model of atmospheric state and dynamics is currently covert and fully intuitive. Data collection, analysis, and Bayesian techniques for modeling and supporting belief updating, and the optimal combination of information from model guidance and TDO awareness of atmospheric evolution and state should be pursued to provide analytical support for performing this challenging task.

3. Although statements are frequently heard about the relative performance of JTWC and CONW guidance, it is important to note that no fully automated (no human-in-the-loop) TC forecasting system exists at the JTWC. As such, there is no sound empirical basis for such statements, at least with the current state of technology at the JTWC.

4. Until the time that a fully or near-fully automated TC forecasting system exists at JTWC, it will prove impossible to determine and quantify the true value contributed by JTWC staff training, knowledge, expertise, and operating procedures to TC forecasting accuracy.

5. The manner in which TC forecasting performance is quantitatively measured, together with human factors research findings in related domains, suggests that forecasting inconsistency, rather than systematic or knowledge-based bias, is likely to be the most significant contributor to the JTWC TC mean track error metric.

6. The manner in which TC forecasting performance is quantitatively measured and communicated (as a scalar value) may provide barriers to learning from experience for TDOs who conceive of TC forecasts in terms of 2 degrees of freedom: speed and direction. A decomposition of overall mean track error into its separable components related to TC speed and directional errors would provide additional support for learning from experience and performance-based feedback.

7. Dedicated efforts by JTWC staff to consistently strive for continuous improvement in addition to performing the mission at hand are truly remarkable in the author’s experience.

1. Introduction

1.1 Background

Navy assets as well as human lives around the Pacific depend on forecasts that the Joint Typhoon Warning Center (JTWC) produces for about 85 tropical cyclones per season with a staff of only nineteen. Making the best use of those limited human resources is essential, and human performance – and therefore forecast accuracy – may currently be limited by sub-optimal design of the task/work environment.

1.2 Scope

The goal of this project was to **evaluate the task and support environments associated with the Joint Typhoon Warning Center (JTWC) watch floor and provide recommendations with the ultimate goal to improve forecast accuracy.**

1.3 Description of Tasks Performed

The contractor has performed the following tasks:

1. Reviewed background documents on the JTWC's mission, responsibilities, and processes in preparation for the site visit. (24 hours)

2. Visited the JTWC during the West Pacific typhoon season in November 2010, to observe the current operations of the JTWC watch floor for a period that depended on the amount of tropical-cyclone activity, but included at least one active tropical cyclone. Six (6) watches were observed in part or whole. (20 hours)

3. Gathered empirical observations of the forecasting routine, available guidance, display products, in-house support efforts provided by JTWC day- working staff, and had discussions with JTWC staff regarding the above as well as the JTWC's feedback and verification processes. Attended In-Brief by CAPT Angove and prepared and presented Out-Brief to CAPT Angove and LCDR Callahan while on site. (15 hours)

4. Compared observations with the state-of-the art in human factors engineering, and made recommendations for potential improvements and additional efforts to improve the forecasters' task environment, to include: information systems, computer models, technological interfaces, and printed text and graphical documents. (16 hours)

5. Analyzed the results of the observational phase and formed hypotheses regarding characteristics of the task environment that may limit forecast accuracy. (15 hours)

6. Formulated recommendations based on the state-of-the-art in human factors engineering from the literature and best practices, and based on empirical

observations. Recommendations include both short-term modifications that could be implemented within JTWC to improve the task environment as well as larger-scale projects that would have potential to yield larger improvements in forecast accuracy. Documented recommendations in a final report delivered to NPS November 16, 2012. (30 hours)

1.4 Level of Effort
Total hours worked: 120

2. JTWC Forecasting Operations

The following sections provide a brief overview of the resources available at the JTWC and the practices used in an attempt to achieve consistently high levels of TC forecasting performance and end-user satisfaction in this challenging work environment.

2.1 Staffing

Nominal JTWC Watch Floor staffing includes 1 TDO (Typhoon Duty Officer), 1 SAT (Satellite Analyst), and 1 TDA (Typhoon Duty Assistant). A second TDO is added for a third, etc., simultaneous storm.

2.2 TDO Warning Cycle Basic Activities

Two cycles of the following nominal activities are performed per 12-hr watch:

1. SAT provides updated fix & intensity estimates via DVORAK
2. TDO sets best track position and intensity based on JTWC and other agency satellite fix positions, satellite imagery, and other observations.
3. Prepare and send BOGUS (input for forecasting models)
4. Create consensus using ATCFS
5. Assess appropriateness of the various models given the larger wx context, known strengths/weaknesses/tendencies of individual models, additional information, etc.
6. Prepare and issue track and intensity warnings
7. Prepare and issue prognostic reasoning summary
8. Handle customer calls, satisfy other requests
9. Be on the lookout for developing storms and create invests

Warning cycle activities can become time-stressed in multi-storm situations, as depicted below in Figure 2-1.

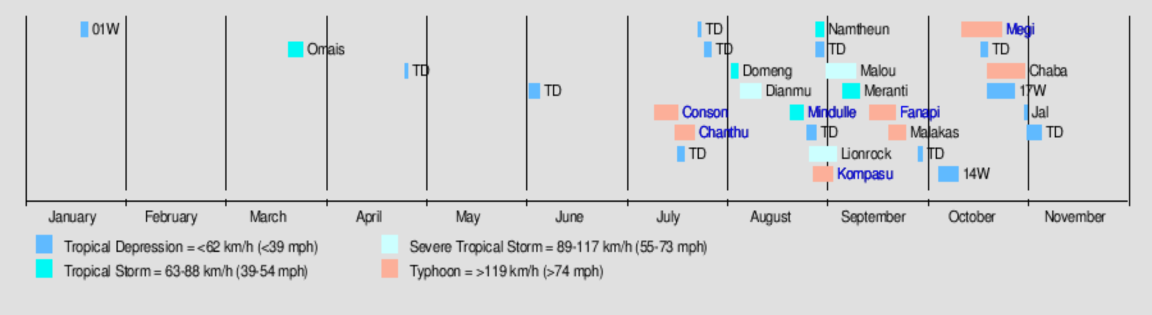

Figure 2-1. *2010 Storm timeline showing significant forecasting overlap intervals.*

2.3 Information Load

During each forecast cycle, the TDO consults, at a minimum, the following information sourses beyond the information available through the ATCF suite:

- WxMap
- NRL TC or FNMOC TC Webpage
- CIMMS TC Page
- Streamlines (Large hardcopy printouts)
- Water Vapor Displays
- CIRA RAMBB
- ASCAT
- JAAWIN U. Wyoming (or other) Collected Observations
- Email (e.g., SATCON Automated DVORAK)
- Microwave Scatterometer
- Agency Analysis Charts

Beyond the simple number of information sources that must be considered, significant barriers to effective information integration across these sources exist. These include the lack of consistent geospatial referencing and scaling, information age (timeliness), source pedigree, data age, reliability or trustworthiness, and so forth. In response to at least some of these concerns, JTWC staff members were observed to have created a prototype information system that would assist them in their information integration tasks. This system, depicted in Figure 2-2, uses Google Earth as a platform on which multiple geospatial layers of displayed information could be presented, toggled on and off, varied for transparency, and so forth.

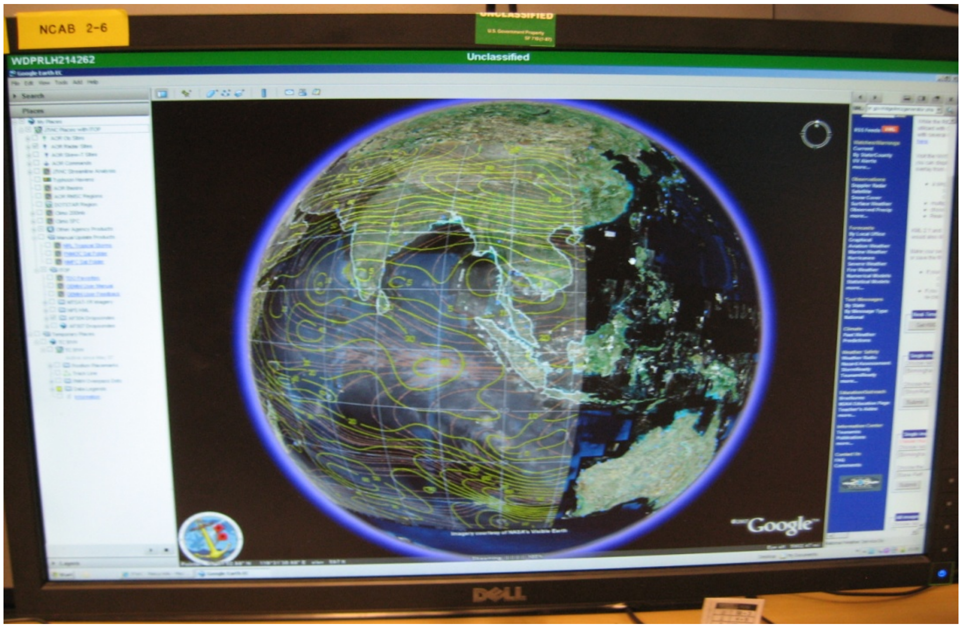

Figure 2-2. *JTWC's Google Earth-based TC forecast information integration aid.*

The Google Earth prototype spontaneously created by JTWC staff sends a strong signal that these staff are overwhelmed by the current cognitive demands they face in attempting to internally (cognitively) integrate the vast array of spatially-oriented information they must consult to continually update their cognitive model of atmospheric evolution. In human factors terms, the Google Earth prototype represents an attempt to create what is known as an integral (rather than separable) display, in which the integration of information occurs within the display itself, allowing the powerful processes of visual perception to process the integrated information directly, instead of relying on the much more fragile and error-prone cognitive processes such as working memory and visual imagery to integrate this information mentally. A standard human factors textbook, such as Wickens, Lee, Liu Becker (2005, see especially chapter 8 on displays) provides ample evidence of the benefits of integral displays to human cognition and performance.

Observations, interviews, and the spontaneous creation of information integration tools such as the Google Earth prototype make it is clear that TDOs continually attempt to update their cognitive model of atmospheric evolution during a watch since the time of the immediately previous model BOGUS initialization. TDOs then use this updated understanding of the atmosphere to influence how they consider and reflect upon subsequent model guidance when it is ultimately received. The following section provides a brief introduction to the role played by model guidance in JTWC TC forecasting operations.

2.4 Model Guidance

The increasingly important role played by numerical-computational models in TC forecasting is well known and documented elsewhere. In recent years, for example, there is a belief that the model consensus known as CONW has approached the performance of JTWC and perhaps even surpassed it over the last few years. Here, the focus is on how the guidance provided by guidance is actually used in the creation of a TC forecast.

Based on observations, interviews, and documents provided by JTWC, the process whereby a TDO and model guidance interact to produce a TC forecast is as follows:

1. Provide bogus to models sent at synoptic times (00, 06, 12, 18 UTC) plus one hour (01, 07, 13, 19 UTC). Bogus data consists of position, intensity (maximum sustained wind speed and minimum sea level pressure), radius of maximum winds, radius of 34 / 50 / 64 kt winds, depth (shallow, medium, or deep), eye diameter (zero if no eye), radius of the outermost closed isobar, pressure of the outermost closed isobar, and 12 to 24 hour past motion (direction and speed movement vectors described previously).

2. The six-hourly models begin their runs at synoptic times (12-hourly models begin at 00 and 12 UTC). During the first hour to hour and a half after synoptic time, the model analysis is constructed from observations (ship, buoy, vertical weather balloon soundings, satellite data, etc.) including JTWC Bogus data (for those models that use this data).

3. The ATCF receives and automatically plots the returned model guidance, for each individual model as well as their overall average track, called CONW. CONW consists of forecast TC positions for T0 + 12, 24, 36, 48, 72, 96, and 120 hours. Due to the delay between model initialization and when their returned guidance is received, an interpolation process is used to update guidance provided by the models, including CONW, into a current forecast, i.e., one consistent with TC evolution since the time of intialization.

4. According to Goerss, Sampson and Gross (2004), this interpolation process operates as follows:

Since forecast track output for the NWP models become available to the forecaster 6 or 12 h after NWP model run time, they arrive too late to be used directly. Instead, the NWP model tracks are interpolated to intermediate times, and then interpolated positions are relocated to reflect the forecaster-analyzed (best track) position. The version of the interpolator used in this study includes a cubic spline (M. DeMaria 2000, personal communication) and a 10-pass, 3-point filter. All interpolated tracks are computed from real-time tracks, not postseason analyzed tracks (best tracks). Quality control for the interpolator includes a linear interpolator to fill in missing 12-

and 36-h forecasts, a forecast position check (the 6-h/12-h old NWP model 6-h/12-h interpolated forecast position must be within 333 km of the current forecaster analyzed position), and a forecast track speed check (60-kt maximum) for all forecast periods beyond 12 h. NWP model interpolated tracks that fail the 12-h forecast position check are eliminated from the interpolator, while those failing the 60-kt speed check are truncated before the 60-kt speed is encountered.

A consensus for a given forecast period is a simple average of the interpolated members that pass the interpolator quality control tests described above. An attempt is made to compute a consensus forecast at the 12-, 24-, 36-, 48-, 72-, 96- and 120-h forecast periods. This consensus is computed if two or more members exist for a given forecast period. If less than two members exist, the consensus is not computed. (p. 634).

5. If the TDO has no reason to adjust or over-ride the guidance provided by CONW, he or she is not able (within the ATCF as currently designed) to directly issue a forecast corresponding to CONW. Instead, it was observed that a non-trivial amount of keypress and mouse work was necessary to create a JTWC forecast corresponding to CONW even if the attempt was to mimic the guidance provided by CONW to the letter. It is understood that the interface could be readily updated to allow the JTWC forecast equivalent to the CONW guidance. This is recommended, *not because it is believed that this will always result in a superior forecast*, but instead to eliminate unwanted, human-induced variability and inconsistency *when the TDO has no desire* to significantly depart from CONW guidance.

6. However, the TDO may in fact have good reasons to over-ride the CONW forecast, given the validity of CONW position and intensity guidance in light of his or her understanding of the current state of the atmosphere. For example, consider the forecast situation depicted in Figure 2-3 below.

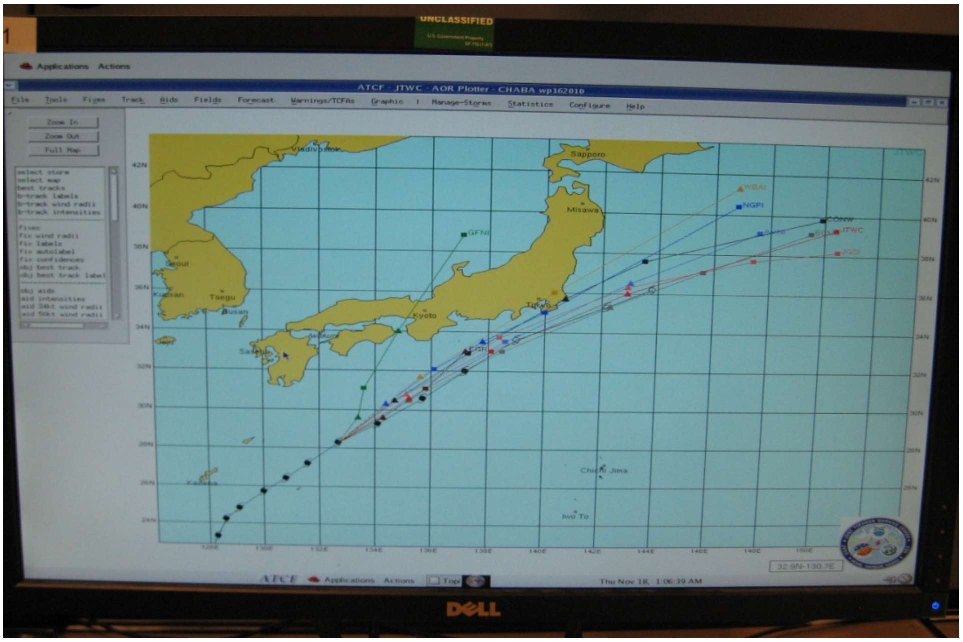

Figure 2-3. *A case in which guidance provided by all but one model (GFNI, to the north) is largely consistent. A TDO commented that the GFNI model was known to diverge northward in this manner due to facts of the particular meteorological situation at hand.*

In the situation depicted in Figure 2-3, the TDO observed that GFNI model was likely to be adding error to the CONW forecast track by drawing it more northward than it should have been (as indicated by the convergence of the rest of the model guidance to the south). In this case, the TDO performed a visual-manual adjustment of the (red) JTWC forecast track a bit to the south of CONW guidance (black), in order to compensate for the erroneous GFNI model guidance.

The type of joint, human-model forecasting behavior described above is consistent with prior research on a technique called the Systematic Approach Forecast Aid (SAFA) used at the JTWC to allow a forecaster to form a selective consensus of model guidance, or SCON (Sampson, Knaff, and Fukada, 2006; also see Carr, Elsberry & Peak, 2001). Experience with SAFA was mixed. For example (from the above):

The Systematic Approach Forecast Aid (SAFA) has been in use at the Joint Typhoon Warning Center since the 2000 western North Pacific season. SAFA is a system designed for determination of erroneous 72-h track forecasts through identification of predefined error mechanisms associated with numerical weather prediction models. A metric for the process is a selective consensus in which model guidance suspected to have 72-h error greater than 300 n mi (1 n mi _ 1.85 km) is first eliminated prior to calculating the average of the remaining model tracks. The resultant selective consensus should then provide improved forecasts over the nonselective consensus. In the 5 yr since its introduction into JTWC operations, forecasters have been unable to produce a selective consensus that provides consistent improved guidance over the nonselective consensus.

Also, the rate at which forecasters exercised the selective consensus option dropped from approximately 45% of all forecasts in 2000 to 3% in 2004.

However, Sampson, Knaff and Fukada (2006) also noted:

To evaluate whether or not the SCON performance in 2001–04 was degraded by relaxing the requirement that the consensus spread be larger than the 250 n mi specified in Carr et al. (2001), statistical analysis for cases in which the consensus spread was larger than 250 n mi are analyzed. For the 30 cases that verified (1.5% of the total number of verifying SCON forecasts), SCON outperformed NCON by about 30%. The results are significant at the 95% level. It is worth noting that the other skillful NWP models and consensus aids available to JTWC forecasters may have influenced the forecasters while performing the SAFA analysis.

It is important to appreciate, however, that these types of forecasting performance comparisons are made somewhat problematic by the fact that the human forecaster clearly adds value to CONW in at least 2 stages of the forecasting cycle: 1) by establishing the BOGUS or some of the initial conditions used by the models; and 2) in the process of using the ATCF to translate the guidance provided by CONW into a timely and feasible forecast. The fact that no *completely automated* system exists for forecast generation (without a TDO "in the loop") currently renders it impossible to determine whether the TDO's activities are adding or subtracting value from what would be achieved by a fully automated forecasting system, that is, one based on CONW integrated with any automatic processing of any other relevant atmospheric information.

2.5 Overall Schematic Model of TC Forecasting Process

The preceding discussion leads to an overall conceptualization or schematic (incomplete) model of the time course of the JTWC TC forecasting process as depicted in Figure 2-4.

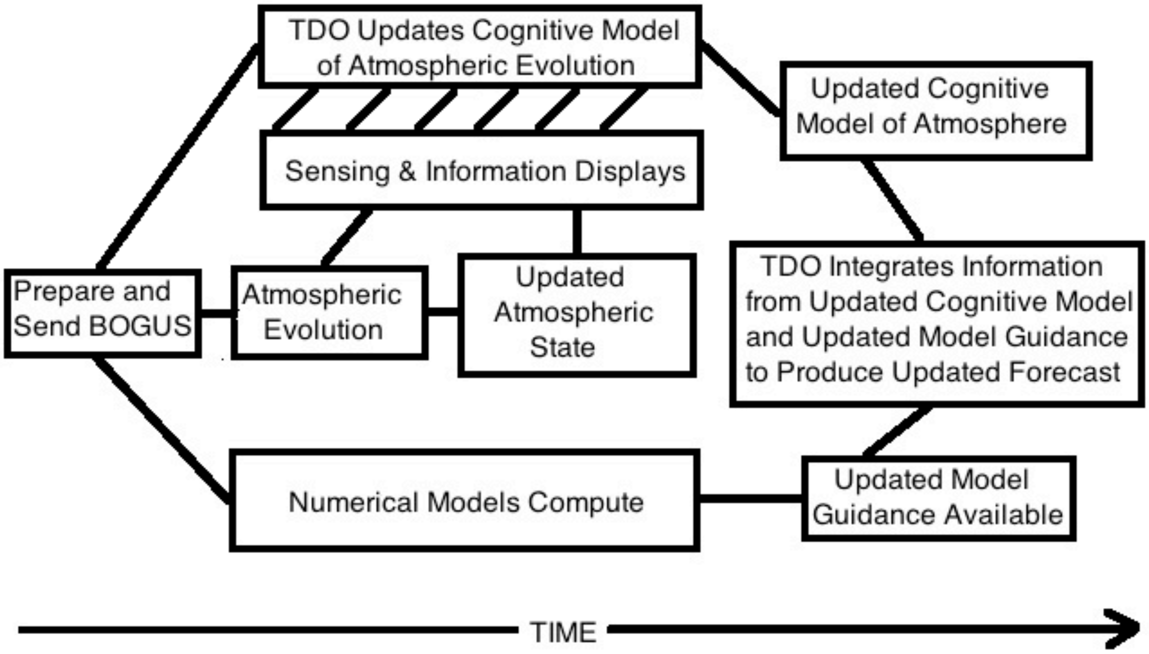
TDO Updates Cognitive Model of Atmospheric Evolution
Updated Cognitive Model of Atmosphere
Sensing & Information Displays
Prepare and Send BOGUS
Atmospheric Evolution
Updated Atmospheric State
TDO Integrates Information from Updated Cognitive Model and Updated Model Guidance to Produce Updated Forecast
Numerical Models Compute
Updated Model Guidance Available
TIME

Figure 2-4. *A schematic model of the JTWC TC forecasting process.*

The model depicted in Figure 2-4 traces the flow of information, cognition, and numerical computation through one forecast cycle from initial model BOGUS-ing to the creation of an updated TC forecast. As depicted in the upper portions of the model, the TDO attempts to maintain "situation awareness" (Kirlik & Strauss, 2006; Strauss & Kirlik, 2006) of those atmospheric dynamics that may affect TC track and intensity over the period of time that numerical models are being run to provide guidance prior to creating and issuing the subsequent TC forecast. The TDO relies on his or her knowledge and expertise in conjunction with real-time information sources identified in section 2.3 above to create a cognitive model of these dynamics. As shown in the large rectangle at the right side of the figure, the ultimate forecast is prepared by the TDO through a process of reflecting on model guidance, such as CONW, in light of this cognitive model of the recent and current atmospheric dynamics and state. On the basis of this process, the TDO may elect to issue a forecast largely aligned with CONW, or may, as discussed previously, decide to adjust the CONW forecast based on atmospheric developments since the time the numerical models were initially BOGUSED, and possibly also information about the tendencies of each of the various models providing guidance.

In terms used in the study of human judgment and decision making, the task faced by the TDO in combining these two sources of information (the information flowing upward from model guidance, and the information flowing downward from the TDO's cognitive model of the atmosphere in Figure 2-4) is termed Bayesian updating or Bayesian belief revision (Edwards, 1962). While obviously much more complex, this task is not unlike cognitive tasks people perform every day in which somewhat outdated information (in the current case, model guidance based on initial conditions known to no longer currently hold) is combined with newly available information (in the current case, updated information about atmospheric evolution since initial model BOGUS-ing) to yield a new belief (in the current case, a TC forecast) that benefits from both sources.

In an everyday context, for example, one might consult today's morning newspaper at lunchtime for weather information to help inform one's decision about whether or not to plan a golf outing later in the afternoon. This forecast has some validity, even though one knows it to be somewhat outdated (e.g., it was produced prior to the production run of the newspaper late the previous evening). Then, one might turn one's attention out the window to scan the skies for signs of rain. This information is certainly more current than the newspaper forecast, but it is far from perfect, and it does not benefit by meteorological information that went into the professional forecast available in the newspaper. What do we know about people's ability to intuitively combine multiple sources of information in an optimal fashion? The following section provides a brief overview of the central findings, with possible implications for JTWC operating procedures or future technological aids.

2.6 Bayesian Analysis of Forecast Belief Updating

There is considerable experimental evidence that people do not update their beliefs optimally (i.e. consistently with Bayes' theorem). This effect was famously documented in experiments outside the subjects' domain expertise by Kahneman and Tversky (1972). This effect has also been documented in experiments within subjects' domain expertise, e.g. in medical professionals interpreting test results (Casscells, Schoenberger, and Graboys, 1978). For a detailed explanation of Bayes' theorem in the context of cognitive engineering, see McCarley and Benjamin (in press).

Recently, however, there have been some successful attempts to design information displays to improve people's ability to reason effectively in Bayesian belief updating. For example, Tsai, Kirlik & Miller (2011) created display aids that were found to significantly improve Bayesian reasoning in the context of intelligence analysis, and Miller, Kirlik, Kosorukoff & Tsai (2008) created display aids that significantly improved the Bayesian reasoning of fantasy sports experts making predictions about professional athletes annual performance in Major League Baseball and National Football league seasons.

While those display aids are not immediately applicable to the JTWC context, prospects for extending these approaches to TC forecasting do exist, especially in the realm of TC intensity forecasting. To do so, historical data would have to be collected (or analyzed, if already available) to determine the base-rate reliability of guidance such as CONW regarding its ability to successfully predict TC strengthening and weakening. Additionally, similar data would have to be collected on TDO's unaided (i.e., prior to viewing model guidance) abilities to successfully forecast TC strengthening and weakening events. With such data in hand, and to return to the previous example on scheduling a golf outing, one would then be in a position to determine how to relatively weight and thus combine the precipitation prediction available from the morning newspaper and one’s intuitive precipitation prediction based on viewing the weather out the window. Similarly, at the JTWC, Bayes theorem could be implemented in computer software as an aid to help a TDO determine how to optimally combine his or her own intuitive prediction of TC strengthening and weakening events, and the guidance obtained about TC strengthening and weakening received from numerical models. Currently, TDO expertise and CONW guidance are combined in an entirely intuitive fashion, without the benefit of analytical support, such as Bayes Theorem, for this challenging and often counter-intuitive task.

3. Analysis of Tropical Cyclone (TC) Forecasting Performance Evaluation

3.1 Why this Analysis is Important

All attempts to improve the performance of a system, whether it be solely human, solely technological, or, as at the JTWC, comprised of both human and technological components, should be based in a detailed understanding of how the performance of that system is ultimately assessed. By contract, the purpose of this report is to make recommendations to improve the performance of the JTWC as an integrated human-technology forecasting system. According to Annual Tropical Cyclone reports prepared by the U.S. Naval Maritime Forecast Center/JTWC, one central component of JTWC performance assessment consists of an annual forecast verification summary.

To conduct this summary, verification of TC warning positions and intensities at 24, 48, and 72-hour forecast lead times are compared against the final best track for each JTWC forecast. Forecast error statistics at both the JTWC and National Hurricane Center (NHC) are computed on the basis of the absolute great circle distance between a forecast position and the corresponding post-analysis best track position. Figure 3-1, after JTWC Annual Tropical Cyclone reports and Tsui and Miller (1988), illustrates nature of the track error calculation for any particular forecast location.

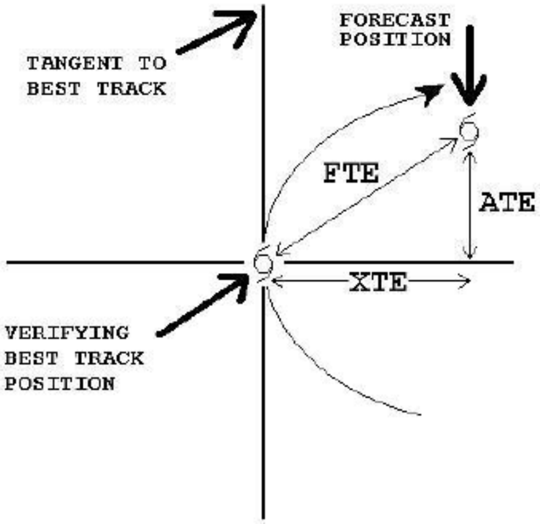

Figure 3-1. *Calculation of track position error (FTE) for a single TC forecast. Annually, these track position errors are summarized in terms of a mean forecast track error. Figure 3-2 depicts gains made over the past years in reducing mean TC track forecast error for the Western North Pacific over the last decades.*

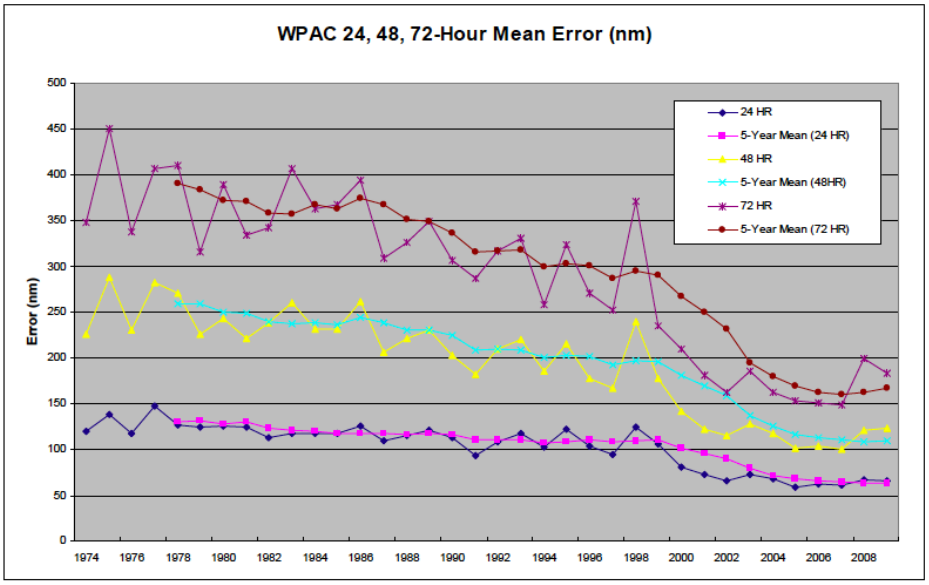

Figure 3-2. *JTWC Mean TC forecast track error between 1974 and 2009 (inclusive).*

Note that the mean JTWC 48-hr forecast track error is now close to the value of the mean 24-hr track error of about a decade ago. These gains are in the same ballpark as those seen in NHC TC forecast track performance in recent years, as depicted in Figure 3-3.

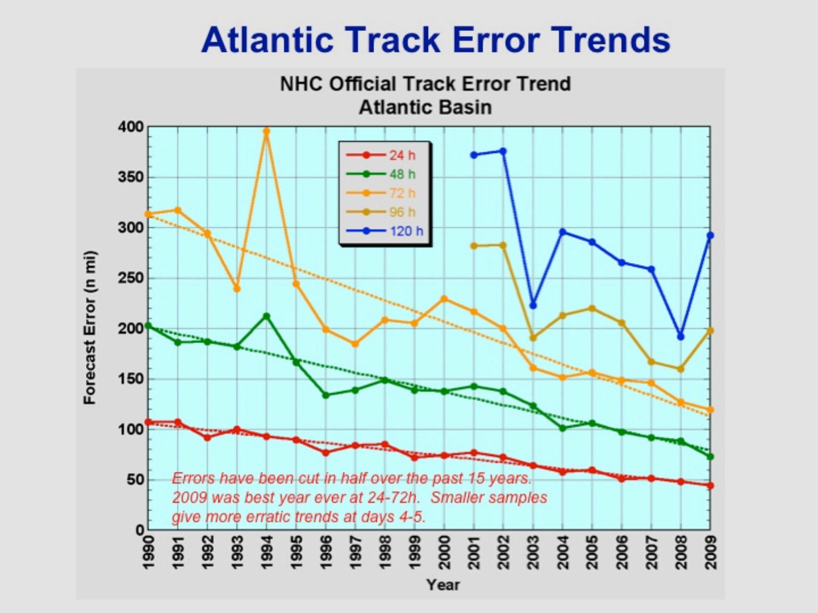

Figure 3-3. *NHC Mean TC forecast track error between 1990 and 2009 (inclusive).*

Close inspection of these (and related) graphs and statistics indicates that NHC mean track forecast errors may be slightly lower than for JTWC in recent years, especially at the 24- and 48-hour lead times. If so, a great many factors may possibly contribute to these differences. Here, the focus is solely on improving JTWC forecast performance based on information collected on-site and a consideration of available theory and best practices in human factors and related areas. No similar on-site visit to NHC was made to inform the conclusions and recommendations of this report.

To this point it seems fairly clear that mean error (typically expressed in nautical miles) plays a central role in the assessment of overall JTWC performance, at least as far as track forecasts are concerned. Intensity forecasts are a more complicated and subtle issue, and as such not a separate focus of the current discussion. However, the ultimate recommendations of this report are intended to serve the general goal of improving JTWC overall TC forecast performance for both track and intensity to the levels that current science and technology allow.

3.2 Minimizing Mean Error: Intuition and Reality

A common and partially correct view of mean error (an average of track error distances over a year or forecasting season) is that there are two components involved: accuracy and precision. Depending on the context, other terms used to describe precision are ‘consistency’ and ‘repeatability.’ These terms (precision, consistency, and repeatability) will be used interchangeably for the purpose of this report. Figure 3-4 provides a graphical depiction of the distinction between accuracy and precision as they exist individually, and how they would collectively combine to produce an overall (scalar) value of mean error.

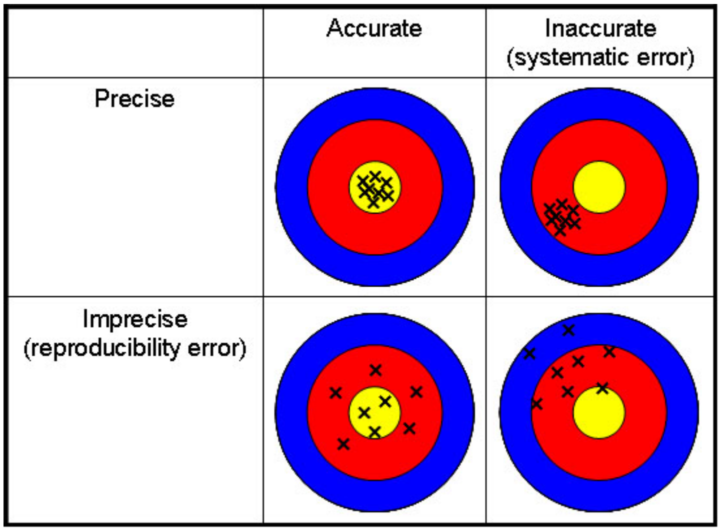
Accurate
Inaccurate (systematic error)
Precise
Imprecise (reproducibility error)

Figure 3-4. *Decomposition of mean (overall) error into separable components of inaccuracy and imprecision. Minimizing mean error (maximizing forecasting performance) requires the highest possible levels of both precision and accuracy. However, when considering Figure 3-4, it is extremely important to note that it leads to false intuitions about the relations between accuracy and precision (or inaccuracy and imprecision) as they combine to determine overall mean error in the TC forecasting context. More specifically, if mean error (of a set of point forecasts, such as the X's in the figure) is the ultimate measure of forecast performance evaluation, then Figure 3-4 leads to the false intuition that a forecaster can be imprecise yet can still be accurate (the lower left quadrant in Figure 3-4).*

However, this is simply not the case, and this fact is likely to create an under-appreciation of just how much even relatively minor levels of imprecision or inconsistency in JTWC track forecasts inflate mean track error beyond what would intuitively seem to be the case.

Why? Note that our intuitive estimation that a set of forecasts or point locations distributed in a pattern like the lower-left quadrant of Figure 3-4 are imprecise yet accurate rests on our intuition that the X's are randomly scattered around the center of the target, and thus the errors "cancel out" in a way that they do not in the lower-right quadrant. It is this "cancelling out" intuition that gives rise to the fundamental idea underlying Figure 3-4 as a whole: that one can demonstrate some inconsistency or inaccuracy yet one can remain, on the whole, largely accurate. While this may hold true of how accuracy is defined in the figure (the central tendency of the distribution should fall in the yellow or target region), it is not at all true of how the mean error of these points would be defined an measured, assuming that they were a distribution of TC forecast locations.

Specifically, recall that a mean (or average value) of a set of numbers is that number that minimizes the sum of the *squared* errors (or squared and therefore the *positive* deviations) from that number.

As such, the error-squared technique results in making all errors contribute positively to the error summing calculation, ruling out any “cancelling out” of positive and negative errors as would be suggested by the lower-left quadrant of Figure 3-4. Thus, the mean error calculation as it is actually implemented violates intuitions that a forecaster can suffer some loss of precision or consistency, yet maintain reasonable accuracy, at the labels in the figure suggest. That is, however intuitive Figure 3-4 may be, the error-squared aggregation method used to evaluate JTWC forecast performance does not favor error patterns that are centered around the best forecast track in the way that Figure 3-4 favors error patterns than are centered around the yellow center of the target (as opposed to those centered elsewhere, as in the bottom-right quadrant of the Figure).

To this point, we have shown that the manner in which JTWC forecasts are assessed:

A) is inconsistent with the intuition that high levels of performance remain achievable by a forecast system that suffers from some level of imprecision as long as the track errors associated with this impression or inconsistency are unbiased (i.e., tend to “cancel out” or center on) the best track location. Instead, as shown in Figure 1, FTE for every forecast location has only non-negative values and thus various errors, such as to the north and south, and to the east and west, do not cancel but sum. Said in other terms, the lay or intuitive understanding of how errors of accuracy and precision combine to yield a scalar value of overall errors does not apply to JTWC performance assessment.

Finally, it is useful to point out a second mismatch between the manner in which JTWC forecast performance is assessed and the manner in which TDOs both conceive of, and communicate with one another, about TC forecasts. Consider the following text that has been excerpted from an official JTWC forecast, in this case, TS/TC/Typhoon 15W (Megi) in October, 2010, including bold highlighting added by this author:

FORECASTS:
12 HRS, VALID AT:
170000Z --- 19.0N 127.7E
MAX SUSTAINED WINDS - 115 KT, GUSTS 140 KT
WIND RADII VALID OVER OPEN WATER ONLY

VECTOR TO 24 HR POSIT: 265 DEG/ 11 KTS
---
24 HRS, VALID AT:
171200Z --- 18.8N 125.4E
MAX SUSTAINED WINDS - 125 KT, GUSTS 150 KT
WIND RADII VALID OVER OPEN WATER ONLY

```
VECTOR TO 36 HR POSIT: 255 DEG/ 11 KTS
                    ---
                    36 HRS, VALID AT:
                    180000Z --- 18.3N 123.2E
                    MAX SUSTAINED WINDS - 135 KT, GUSTS 165 KT
                    WIND RADII VALID OVER OPEN WATER ONLY
VECTOR TO 48 HR POSIT: 255 DEG/ 10 KTS
                    ---
                    EXTENDED OUTLOOK:
                    48 HRS, VALID AT:
                    181200Z --- 17.8N 121.1E
                    MAX SUSTAINED WINDS - 090 KT, GUSTS 110 KT
                    WIND RADII VALID OVER OPEN WATER ONLY
                    VECTOR TO 72 HR POSIT: 270 DEG/ 08 KTS
                    ---
                    72 HRS, VALID AT:
                    191200Z --- 17.7N 117.7E
                    MAX SUSTAINED WINDS - 100 KT, GUSTS 125 KT
                    WIND RADII VALID OVER OPEN WATER ONLY
                    VECTOR TO 96 HR POSIT: 280 DEG/ 06 KTS
```

The highlighted elements of this forecast, fully consistent with observations of, and discussions with TDOs at JTWC, indicate that, by and large, forecasters think of TCs as dynamic storm systems moving with a particular speed and direction (DEG/KTS) at any point in time. That is, TDOs tend to think of TCs in polar (rather than Cartesian) coordinates, where motion is defined by a vector whose angle indicates direction of movement and whose length indicates speed of movement. This is only natural, as this is how the evolution of a TC is actually experienced.

Consider, though, the difficulty of learning for a forecaster whose job it is to make predictions of TC speed and direction yet is ultimately assessed on mean track error. An analogy would be to a golfer who, after a session practicing putting, was provided feedback that his or her putts ended, on average, at a distance 2.4 meters from the hole. What the golfer needs is more diagnostic feedback: what does this 2.4 meter measure mean when it comes to what is actually being controlled: aim (or putting direction) and distance (or speed)?

The TC forecaster, like a golf putter, lives and thinks in a world of speeds and directions, yet the TC forecaster, and the JTWC as a whole, is evaluated by a scalar metric of track error, measured in a Euclidean rather than Cartesian coordinate system. As such, it is useful to consider the mapping between errors in forecast TC speed and direction (individually) and overall mean track errors. Here, too, it may be that this mapping (which, in theory at least, needs to be understood for a forecaster to learn to adjust his or her speed and directional forecasts on the basis of overall track error) may not be fully intuitive. If not, it may pay to pursue the development of aiding or training technologies that provide TDOs with a better understanding of these relationships. Figure 3-5 provides a starting point for doing so.

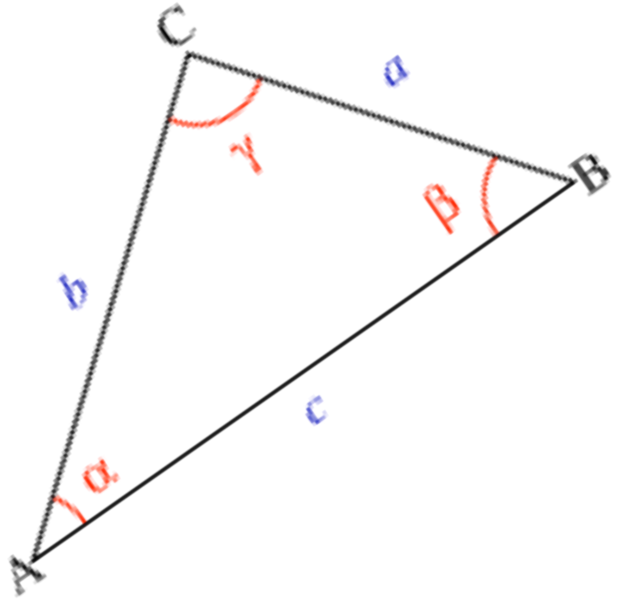

Figure 3-5. *A TC forecast situation depicted to enable a convenient transformation between polar and Cartesian coordinate systems. Current TC location is vertex A. Forecast TC location (in 24, 48, etc.) hours is vertex B, at a distance c, or in terms of TC movement at a speed of c units of distance per unit time. Best track (actual) TC location at this forecast lead-time is vertex C, at a distance b, or in terms of TC movement at a speed of b units per unit time. The angular or directional error of the forecast is given by a. The speed error of the forecast is (c – b) in absolute terms, or the ratio (|c – b | / b) in relative terms. The Forecast Track Error (cf. Figure 2-1) is given by the length of side a.*

Based on observations and discussions at JTWC, evidence indicates that TDOs decompose the task of achieving a low FTE into two components, typically with strong reliance on model guidance: first getting the storm track direction (or angle) correct, and then getting the storm speed correct. Figure 3-6 can be used to better understand how these two components of TC forecasting combine to determine an ultimate FTE, or an average track error over the course of a storm or a season. A consideration of the geometry depicted in Figure 3-6 (see caption for explanation) and the law of cosines yields Equation 3-1

Equation 3-1. $$a^2 = b^2 + c^2 - 2bc\cos\alpha$$

Equation 3-1 makes it clear that there is a non-linear (and thus, likely to be non-intuitive) relationship between accuracy in forecasting TC speed (achieving *c* = *b*), accuracy in forecasting track direction (achieving a = zero), and achieving overall accuracy as measured by FTE (achieving a value of *a* = zero). Figure 3-6 depicts this non-linear relationship.

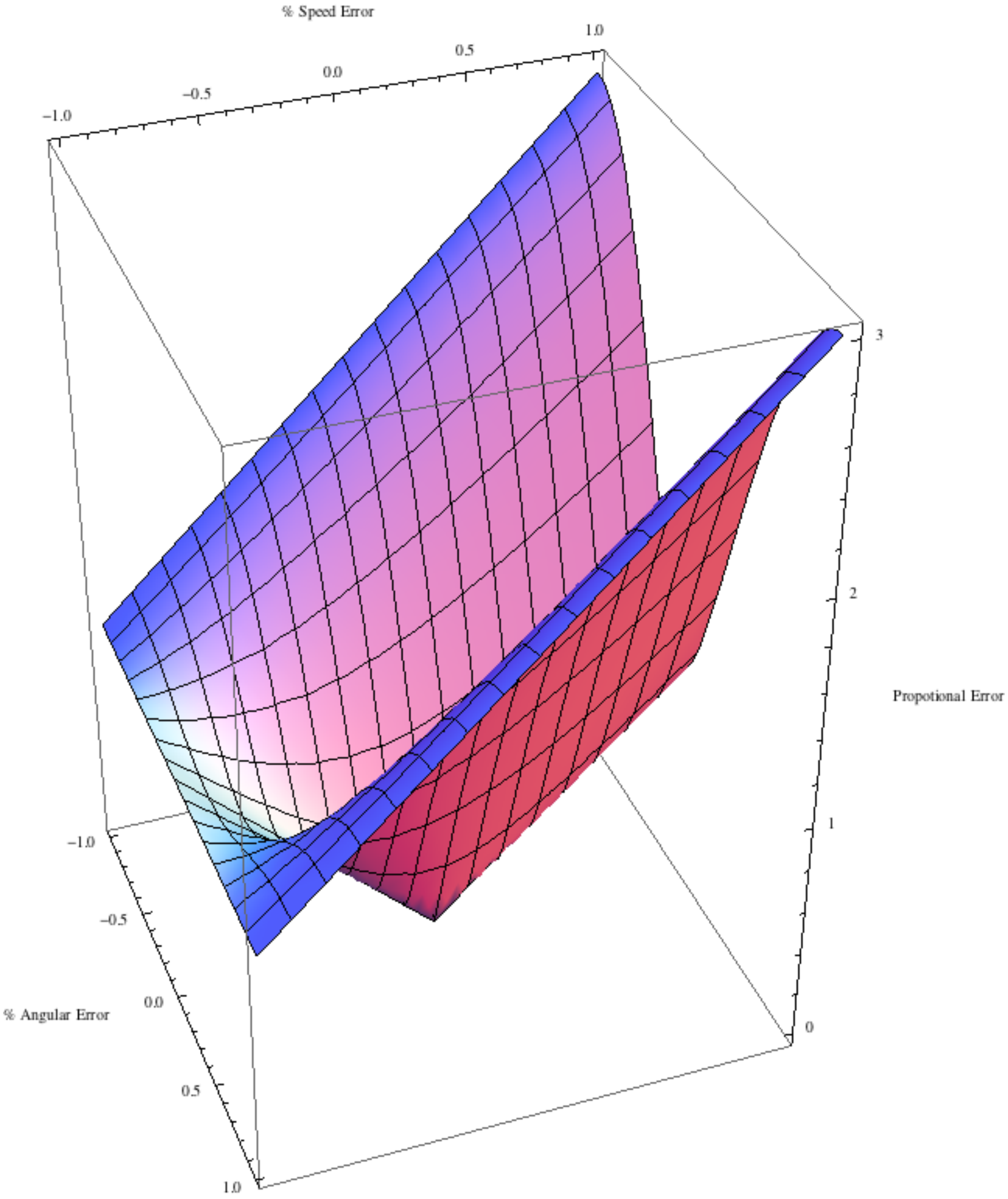

Figure 3-6. *Forecast Track Error (labeled "Proportional Error" in the figure) shown as a function of TC track angular (or direction) error and TC speed error. All errors are given as proportional (or %) measures to depict the general case (thus, "Proportional*

Error" is given rather than FTE in any particular unit of distance). The central "bottom" point in the graph (% Speed and % Angular errors = 0.0) results in a value of Proportional Error of 0.0, lying on the "floor" of the 3-D graph.

Figure 3-6 is useful for obtaining an immediate appreciation of the non-linear relationship between speed and directional errors and overall FTE in TC forecasting, yet it is difficult to obtain a precise understanding of the tradeoffs involved. Figure 3-7 below is a more useful graphical representation for this purpose.

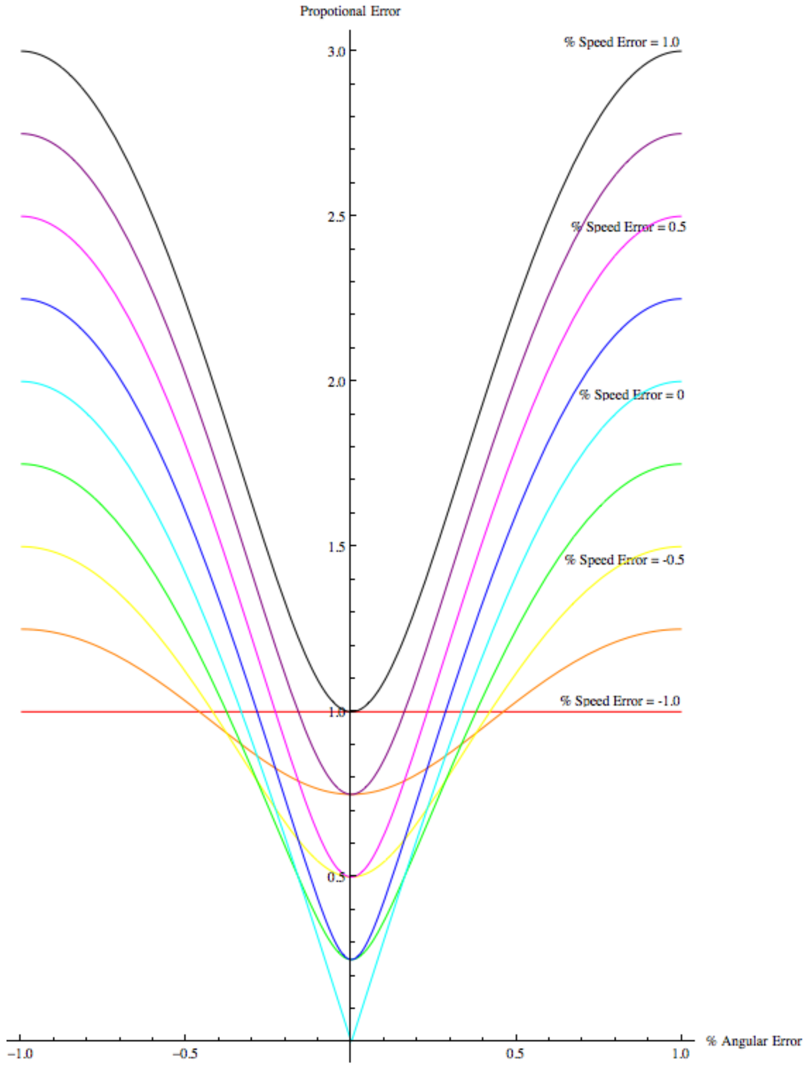

Figure 3-7. *A 2-dimensional depiction of the contributions of forecast TC speed and angular (direction) errors on Proportional Error (or FTE).*

In both Figures 3-7 and 3-8, speed and angular errors are to be interpreted as follows. A speed error of 1.0 is an error of forecasting TC track speed to be twice its actual value, a speed error of 0.0 is forecasting track speed equal to its actual value, and a speed error of

-1.0 is an error of forecasting track speed to be 0.0 when the storm is actually moving. Angular error is 0.0 when the forecast track direction is correct, is 0.5 when the forecast is 90-degrees to the right of the best track, and -0.5 when the forecast is 90-degrees to the left of the best track.
Figure 3-8 below summarizes the gist of this analysis of the combination of TC speed and directional errors on overall TC forecast track error (FTE). Each colored region depicts a region of approximately constant FTE. The current location of the TC is assumed to be at the lower center point in the diagram: Coordinates (0.0, -1.0). The next (known, or best track) location of the TC for the next forecast period is assumed to be at the center (0.0, 0.0) point of the diagram.

Note that, intuitively, this graph is symmetrical about the Y-axis, that is, for errors in either the rightward or leftward track directions. What is not so immediately intuitive is that the graph is not nearly symmetrical about the X-axis, that is, for errors underestimating TC speed (the lower half of the graph) and for those overestimating speed (the upper half of the graph).

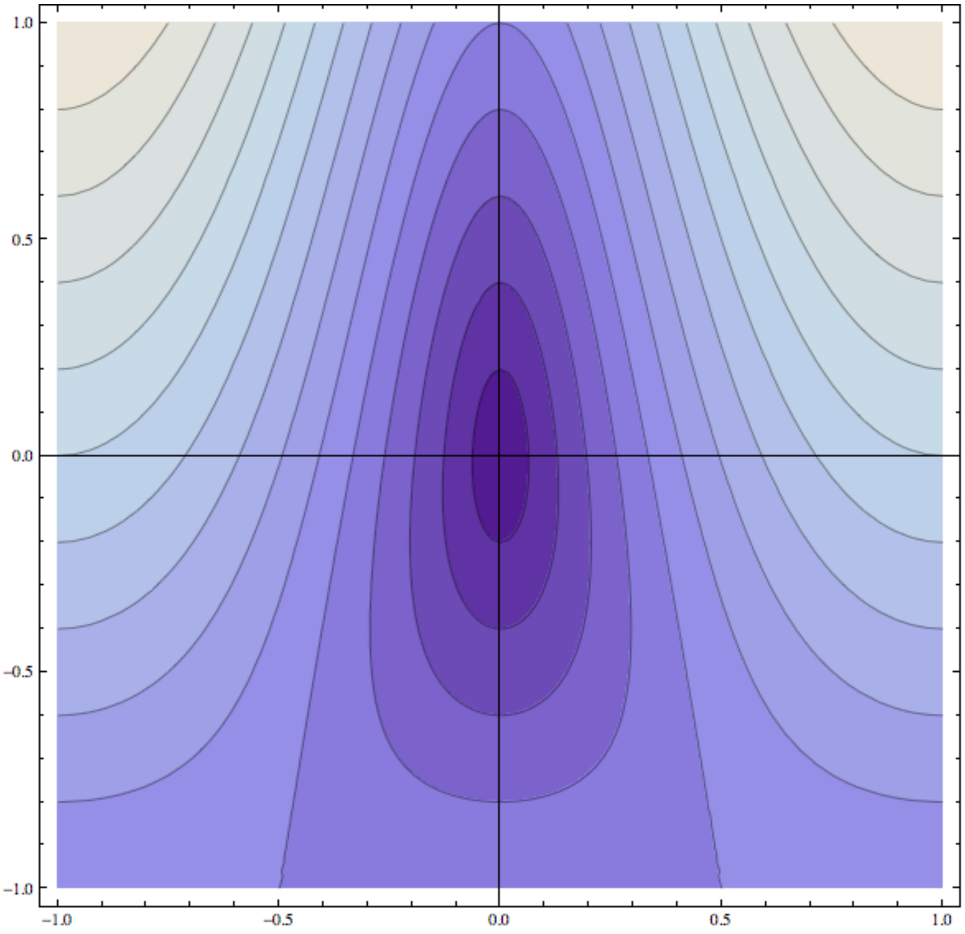

Figure 3-8. *Iso-FTE-Error regions in TC track forecasting as a function of errors in forecasting track direction or angle (X-axis), and errors in forecasting track speed (Y-axis). The current TC location (at the time of the forecast) is at location (0.0, -1.0) or the lower-center of the diagram, and the true (best track) location at the forecast lead-time is at the center of the graph (0.0, 0.0).*

Figure 3-9 can be read as a mapping of a non-linearly shaped "penalty function" for errors in forecasting TC speed and direction, clearly showing that these errors combine in complex ways to determine ultimate FTE. During observations and interviews at JTWC, it was evident that TDOs often have strong intuitions about the behavior of TCs they are forecasting, and even when model guidance (either model consensus, or the guidance of a particular model) should be followed and when it should be not. However, no procedural, training, or technological guidance was observed that would provide TDOs with complementary intuitions about how (relative) errors in both their TC speed and direction forecasts would be penalized by the overall mean track error statistics that would ultimately be used to assess JTWC's forecasting performance.

For example, it is readily seen from Figure 3-8 above that an angular or directional error of a particular degree or size will be penalized more heavily in the assessment of mean track error when the speed of a TC is high, as compared to when a TC is moving more slowly. This may have implications for how a TDO might best "hedge his bets" in times of high forecast track uncertainty, and perhaps even for a consideration of whether mean forecast track errors may be somewhat artificially inflated during periods when TC speeds are greatest during dissipation and poleward curvature over open oceans.

3.3 Summary

To this point, we have shown that the manner in which JTWC forecasts are assessed:

A) is inconsistent with the intuition that high levels of performance remain achievable by a forecast system that suffers from some level of imprecision as long as the track errors associated with this impression or inconsistency are unbiased (i.e., tend to "cancel out" or center on) the best track location. Instead, as shown in Figure 1, FTE for every forecast location has only non-negative values and thus various errors, such as to the north and south, and to the east and west, do not cancel but sum. Said in other terms, the lay or intuitive understanding of how errors of accuracy and precision combine to yield a scalar value of overall errors does not apply to JTWC performance assessment.

B) is inconsistent with the assumption that high levels of performance remain achievable by a forecast system that may be unaware of non-linearity in the manner in which forecasts of TC speed and direction combine to determine overall mean forecast track error. It is just as natural and expected that TDOs decompose the TC track forecasting task into individual components of direction and speed as it is to note that golf putters (and golf putting instruction) does so. Given the non-linearity involved, without additional training or technological support, it is too much to ask for forecasters to determine how to use a unified, scalar measure of track error to learn to adjust his or her separate forecasts of both TC speed and direction.

It has been established that TDOs at JTWC work in an environment that is extremely unforgiving of error, and especially (and possibly even counter-intuitively) of errors associated with inconsistency. In addition, the mean track error statistic used to evaluate JTWC forecasting performance may not be fully intuitive in the manner in which it excessively punishes even a very infrequent number of large errors against a background of a very high number of highly accurate forecasts. Finally, this scalar mean track error measure, considered as potential feedback from which to learn, is not particularly diagnostic and focused with respect to how TDOs conceive of TC behavior and their prediction, namely, in terms of movements with a particular speed and direction.

4. A Baseline for Comparison in Naval Operations

We bring to this study of TC forecasting at the JTWC some previous experience in conducting research on time-stressed judgment in US Naval operations. In 1988, the USS *Vincennes* mistakenly shot down an Iran Air commercial jetliner over the Persian Gulf. As a result, the U.S. Office of Naval Research established a research program on Tactical Decision Making Under Stress, or TADMUS.

Our TADMUS research was one of the many efforts initiated and supported under the overall program (for a comprehensive account of TADMUS research see Cannon-Bowers & Salas, 1998). As described in a chapter written with our colleagues in that volume (Kirlik, Fisk, Walker, & Rothrock, 1998; also see Bisantz, Kirlik, Gay, Phipps, Walker & Fisk, 2000), one of the initial steps in our own research was to visit a naval pre-commissioning team training site, consisting of a full-scale hardware and software simulation of a ship-based Combat Information Center (CIC). At this site entire CIC teams received tactical decision-making and crew coordination training just prior to taking to sea and conducting active operations. We focused our observations and subsequent research on the task of the Anti-Air Warfare Coordinator (AAWC), who was responsible for using a computer workstation containing a radar display and a wide variety of other information sources to make identification judgments of initially unknown objects, called "tracks," in the environment of his or her ship.

Through field observations, interviewing performers and trainers, preliminary task analysis and a review of the literature, we determined that the central task of the AAWC was largely consistent with the image of judgment portrayed by Brunswik's lens model (Brunswik, 1955; also see Kirlik, 2006 for a wide variety of applications of the lens model in technological systems and workplaces). A primary challenge faced by this performer was to use multiple, locally (or "proximally") displayed information sources (or "cues") of various degrees of reliability (or "ecological validity") in order to identify remote (or "distal") environmental objects, in this case, the identities and properties of tracks presented on a radar display. Modeling cognition in this case, or more specifically, judgment under uncertainty, is the focus and purpose lens modeling, to which we will return in a following section.

But at this level of description, the parallels between the task of the AAWC performing in a CIC filled with graphical and numerically displayed information and a TDO performing at the ATWC should be evident. Both are faced with the difficult challenge of using a wide variety of information sources of varying levels of validity to try to infer the behavior of remote systems having a only a limited degree of predictability. Successful performance in such cognitively challenging tasks is known to be depend on having the knowledge required to perform at a high level, as well as the ability to execute one's information processing using that knowledge in a highly consistent fashion.

4.1 Knowledge versus Execution

A common distinction is made between two factors limiting performance in cognitive tasks, and especially those performed in dynamic environments under time stress: the content knowledge required for high levels of accuracy, and the efficiency or consistency with which knowledge-based performance strategies are executed. Ideally, technological design and training interventions should selectively address whether observed human performance limitations are due primarily to deficits in task knowledge, or instead due primarily to the inability to apply what is known to the task at hand in highly consistent or repeatable fashion. Of course, it may be the case in any particular instance that both of these factors may play a role.

To illustrate this distinction, which plays a fundamental role in the analysis and modeling that follows, consider the task of performing a series of 100, 2-digit multiplication problems (e.g., 12 x 35, 91 x 11) using pencil and paper. Imagine that two groups of subjects in a laboratory experiment were asked to perform this task: elementary school children who had only recently learned multiplication, and college seniors. Also imagine that each group was asked to perform this task under two conditions: in the first, one hour was allotted for the task, while in the second, only 5 minutes was allotted. What results would we expect to see? For the purpose of this example, assume that task knowledge corresponds to knowledge of how to multiply any two single digits along with a strategy for correctly performing 2-digit multiplication (carrying, and so forth). What might one expect the results of such an experiment be?

First, one would (hopefully) expect that college seniors would perform this task nearly flawlessly when given one hour to do so, indicating that they not only possessed the necessary task knowledge, but were also able to execute their knowledge-based strategies in a nearly flawless manner. In contrast, we may expect a higher proportion of errors in the elementary student group. A detailed analysis of these latter errors could be used to reveal whether, for any particular student, the errors were systematic in some fashion (e.g., forgetting to carry, erroneous knowledge of what 7 times X equals, and so on). Any such systematic errors would signal knowledge deficits, as opposed to error patterns appearing unsystematic or random (e.g., a child who got the same exact problem right twice but erred the third time), which would more likely signal deficits of execution.

Now consider the results one might expect to see in the 5-minute condition. Almost certainly, we would expect both groups to make more errors. But consider the college student group. Does this result indicate that they somehow lost some of their knowledge of multiplication? Unlikely. Instead, it is more likely that their errors would be largely non-systematic; i.e., slips rather than mistakes, because they had previously demonstrated complete knowledge required for the multiplication task.

Turning back now to the dynamic, time-stressed CIC and JTWC contexts and the AAWC and TDO judgment tasks, important implications for technology design, operating procedures, and training could result if we could tease apart whether any observed performance limitations were due largely to knowledge deficits, or instead to failures to successfully execute judgment strategies using this knowledge.. As demonstrated in the following, the lens model is useful in this regard since it provides a means for decomposing judgment performance into the quality of a performer's task-relevant knowledge and the quality of a performer's ability to make consistent judgments on the basis of this knowledge. This may be an especially useful distinction to be able to make when studying judgment under conditions of high information load, time stress, and uncertainty characteristics of many technological, operational contexts such as the JTWC.

In the context of the lens model, "knowledge" is taken to mean knowing which of the many candidate judgment cues or information sources are useful, their relative reliability or ecological validity, and how they should be weighted and combined to arrive at a judgment. It is also important to mention that lens model analysis also allows one to diagnose the extent to which observed performance limitations are not due to knowledge or execution limitations on the part of the human at all, but are instead due to inherent uncertainty in the performer's task environment. In such cases, training is insufficient to improve judgment performance. Here, performance can be improved only by enhancing the overall reliability of the proximally displayed information (e.g., by improving or adding sensor or display technology, the development and implementation of superior numerical TC models to provide improved guidance, etc.).

4.2 CIC Modeling and Results: Implications for the JTWC

Figure 4-1 provides a graphical depiction of the lens model created specifically for describing AAWC performance in the CIC track identification task. The left side of the figure depicts the model of the environment, which describes the relationship between the judgment criterion value (e.g., friendly, hostile, commercial airliner) and the cue values available at the time a judgment was made. The model of the human, shown on the right side of Figure 4-1 represents the relationship between the cue values and a participant's judgments (the *judged* criterion value), and represents the participant's policy or strategy. In these two models, the actual criterion value and the judged criterion value are represented as linear combinations of the cue values. Thus, these two models are linear regression models of participant judgments and the environmental criterion.

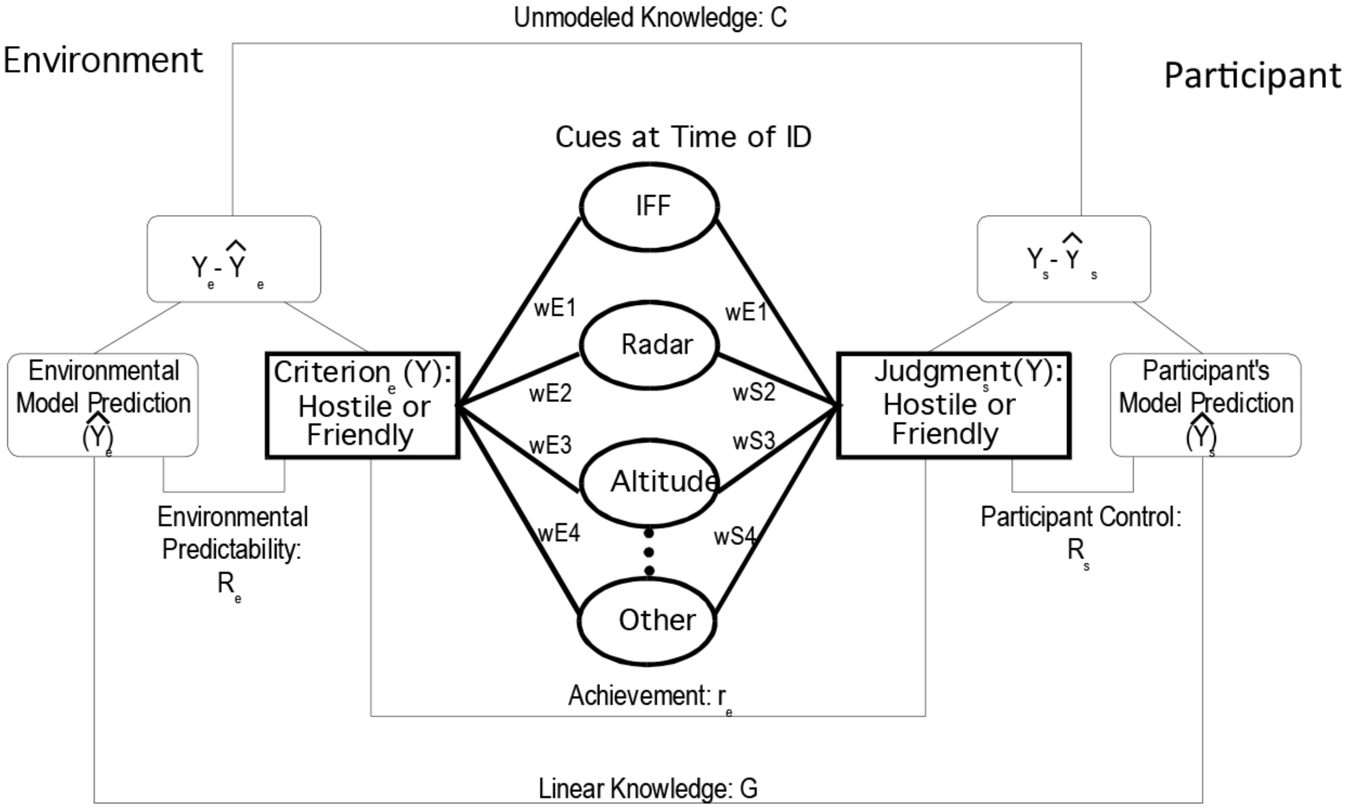
Unmodeled Knowledge: C
Environment
Participant
Cues at Time of ID
IFF
Radar
Altitude
Other
$Y_e - \hat{Y}_e$
$Y_s - \hat{Y}_s$
wE1
wE1
wE2
wS2
wE3
wS3
wE4
wS4
Environmental Model Prediction $(\hat{Y}_e)$
Criterion$_e$ (Y): Hostile or Friendly
Judgment$_s$(Y): Hostile or Friendly
Participant's Model Prediction $(\hat{Y}_s)$
Environmental Predictability: R_e
Participant Control: R_s
Achievement: r_e
Linear Knowledge: G

Figure 4-1. *Lens model depiction of CIC track identification task as performed by an AAWC.*

By comparing aspects of these two linear models, the relationship between human judgment policies and the structure of the environment can be described. This comparison is performed using the lens model equation (LME):

$$r_a = GR_S R_E + C\sqrt{1-R_S^2}\sqrt{1-R_E^2}$$

In this equation, r_a, known as achievement, measures how well human judgments (predictions, forecasts, etc.) correspond to the actual values of the environmental criterion to be judged. Achievement is shown in Figure 4-1 as a overarching curved line at the top of the figure linking judgments to criterion values. In the track identification task, achievement corresponds to how well participants judged the actual identity of the track. This measure (r_a) is calculated as the bivariate correlation between the participants' judgments and the values of the (actual) environmental criterion. In a TC forecasting task at the JTWC, achievement would analogously be measured in terms of bivariate correlations between forecasted storm properties (e.g., intensity, location, etc) at particular lead-times, and their "ground truth" values as determined by post-hoc analyses of best track.

Lens model parameter *G*, often called knowledge, measures how well the predictions of the *model* of the human judge (again, "predictor," "forecaster," etc.) match predictions of the *model* of the environment. *G*, shown in Figure 4-1 as a line linking predicted judgments to predicted criterion values, measures how well the linear model of the judge matches the linear model of the environment: if the models are similar, they will make the same predictions. Thus, it reflects how well a modeled judgment policy captures the linear structure in the environment, and can be seen as measuring a judge's knowledge of, or adaptation to the environment's cue-criterion structure. *G* is calculated as the correlation between the *predictions* of the participant model (predicted judgments) given a set of cue values, and the *predictions* of the environmental model (predicted criterion value) given the same set of cue values.

$R_{e,}$ shown as a line in Figure 4-1 linking actual to estimated criterion values, measures how well the value of the environmental criterion can be predicted with a linear model of the cues. That is, it measures the adequacy of a linear model of the environment. Thus, the value of R_e for the track identification task measures how linearly predictable a track's identity was given its cue values, and is considered a measure of environmental predictability. R_e is computed as the correlation between the predictions of a linear model of the environment (e.g., predicted criterion values) given a set of cue values, and the actual criterion values.

$R_{s,}$ shown in Figure 4-1 linking human to predicted judgments, is a parallel measure to $\mathrm{R_e}$ and provides an estimate of how well human judgments can be predicted with a linear combination of the cue values. For the CIC track identification task, $\mathrm{R_s}$ described how well an AAWC's identifications of a track could be predicted given a linear combination of the track's cue values. Higher $\mathrm{R_s}$ indicates that an AAWC made judgments more consistently with respect to a linear model. $\mathrm{R_s}$ is considered a measure of cognitive control or the consistency with which a judgment strategy is executed by a performer. If a performer's behavior is not well predicted by a linear model of the participant's own judgments in tasks where evidence exists that a linear-additive model should be descriptive, then a low value for $\mathrm{R_s}$ suggests that a performer is not consistently executing the strategy represented by that model. $\mathrm{R_s}$ is computed as the correlation between the outputs of a linear model of a performer (e.g., predicted judgments) given a set of cue values, and the actual judgment.

Finally, C, shown in Figure 4-1 linking the differences between predicted and judged or actual criterion values, measures the extent to which partipant's judgment strategy and the environmental structure share the same unmodeled (in this case, nonlinear) components. Nonlinear cue usage is nearly always found to be negligible.

The lens model equation indicates that each of these factors (environmental predictability: R_e, cognitive control or consistency of strategy execution: R_S, and modeled knowledge: *G* contributes to overall task achievement (r_a), and each of these factors can be *individually* estimated in analyzing judgment performance.

Data on track identification judgments from a simulation of the CIC task context were analyzed using the lens model, with results as shown in Figure 4-2 below.

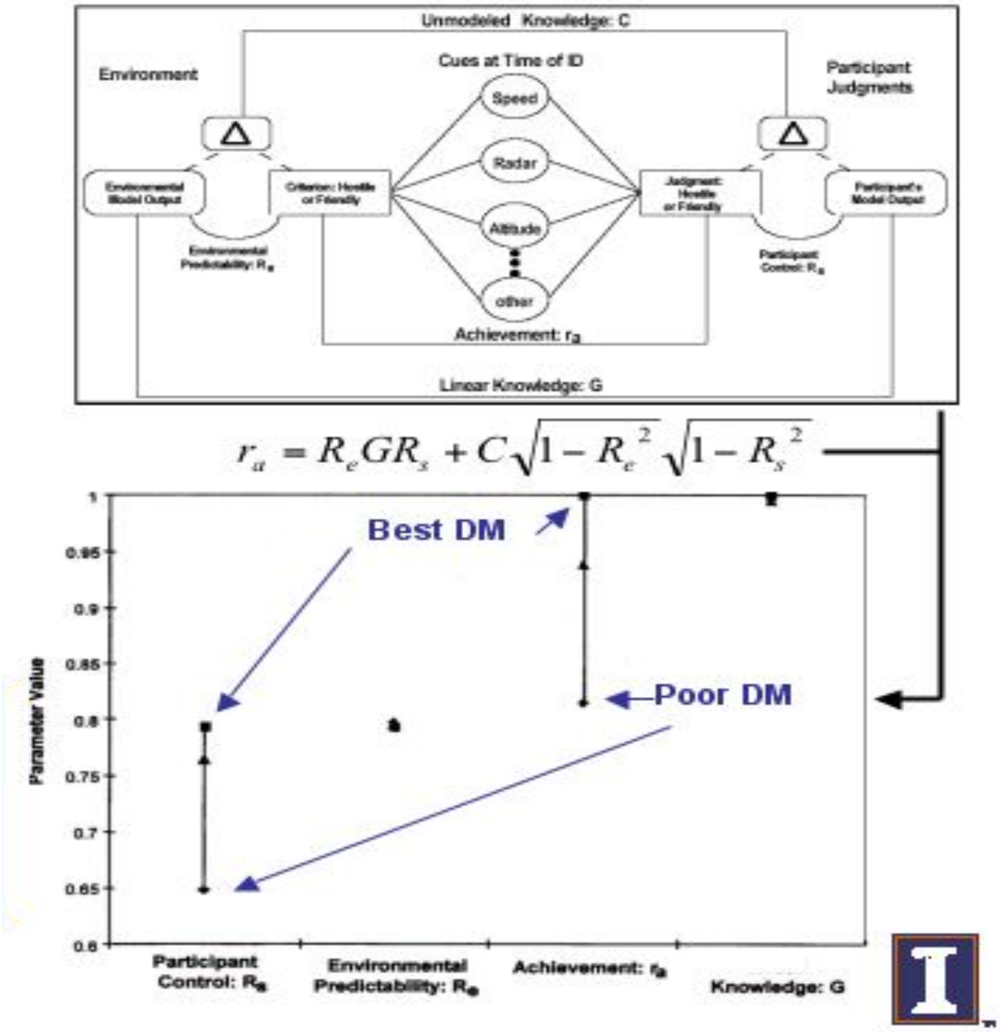

Figure 4-2. *Lens model of the CIC case, and results showing that the primary factor distinguishing high from low achievement ("DM" for decision making) was cognitive consistency (or "participant control"), rather than task knowledge (or G).*

Although time did not permit an analogous lens model analysis of the JTWC context and TDO forecasting performance, there are no reasons to believe that the results of this prior CIC research would not apply to the JTWC, as it is a similarly information-rich, yet uncertain and time-stressed work environment. TDOs are knowledgeable about their tasks, yet, aside from the ATCF itself, have insufficient resources (and in multi-storm situations, perhaps even insufficient time), for *consistently* navigating among the highly diverse (in both content and form) sources of information available to them, for integrating this information cognitively (thus the spontaneous creation by JTWC staff of a Google Earth tool for integrating this information externally), nor for consistently implementing the results of cognitive information processing into forecast products.

5. Conclusions and Recommendations

Based on a review of available documents, information gathered at the JTWC through observations, interviews, discussions, surveys of information technology and interfaces on the watch floor, a variety of mathematical analyses and related research findings presented in earlier sections of this report, the best available theory of human judgment (forecasting) under uncertainty, and best practices in human factors engineering, the following conclusions and recommendations are provided:

5-1. TC forecasters at the JTWC operate in an environment that rewards both accuracy and consistency or repeatability. While the accuracy dimension is highly intuitive, the manner in which inconsistency contributes to inflating mean error is less intuitive:

5-1-1. Every TDO may not fully appreciate that, unlike the lay notion of imprecision caused by inconsistency, errors of positive and negative sign (north versus south, east versus west, slow versus fast storm motion, under- versus over-estimation of intensity) do not cancel out, but instead always sum.

5-2. TC forecasters at the JTWC have a natural tendency to decompose track forecasts into two dimensions: movement direction and speed. As such, the intuitive coordinate system in which TDOs conceive of, and predict, storm movements is inherently polar, rather than Cartesian. Mean track error is calculated in terms of Euclidean distance in a Cartesian coordinate system. As such, there is a mismatch between the way these professionals conceive of the task, and how they are evaluated. Analyses have been provided that demonstrate that the nature of this mismatch is inherently non-linear, and thus not likely to be intuitive to every TDO. This compromises the ability of a TDO to learn most effectively from mean track error feedback ("how much was due to misjudging speed?" - how much direction or curvature?"), and compromises the TDOs ability to make track forecasts that are sensitive to the non-linear penalty function that maps storm speed and direction errors into a scalar value of mean track error. A variety of training or interface design interventions could address this issue.

5-3. Ample evidence from a variety of directions all points to the conclusion that inconsistency is the primary factor limiting JTWC forecast performance as assessed in terms of mean track error. Inconsistency can be mitigated by:

5-3-1. A fully integrated suite of information displays with common geo-referencing, scaling, syntax, etc. Ideally, displays should carry a common coding scheme for the age of data, and, if measurable, the validity or certainty level of information.

5-3-2. Standard operating procedures that are more precise in terms of exactly which information sources should be consulted in which contexts, in which order, along with associated checklists.

5-3-3. Minimizing non-mission critical interruptions of the TDO on the JTWC watch floor.

5-4. Technology to allow the TDO to automatically issue CONW forecasts is needed. The question of how JTWC's performance compares with CONW cannot be definitively answered until this feature is available, due to the current level of TDO participation in the construction of a CONW forecast, even if the TDO has no reason or desire to over-ride CONW.

5-5. With the above technology in place, it will then become possible to empirically determine the degree to which additional TDO intervention adds or subtracts value from CONW, and in which situations. Various experiments can be conducted, regardless of which (CON, SCON, manual) of various forecasts are formally issued by JTWC. These data should be subjected to additional analysis and modeling.

5-6. TDOs currently combine information from their expertise and intuition with guidance provided by numerical models in a largely covert, unsupported, and purely intuitive fashion. This information integration task can be formulated as an optimal belief updating task, using the formalism of Bayes Theorem. If data were collected so that the various quantities necessary to implement Bayes theorem were available, then the belief updating and information integration process could be aided with the design and implementation of a software tool to analytically support this challenging aspect of the TC forecasting task. Intensity forecasts would provide an especially attractive arena for testing this approach, as hypotheses about intensity changes are more readily endurable (increase, decrease, no change) than are hypotheses about future track.

5-7. Mean track error is a deterministic assessment of what is known to be an inherently probabilistic (uncertain) forecast, whether all current forecast products are formatted in these terms or not. Alternative forecast products for better communicating forecast track uncertainty should be explored. Alternative measures of assessment of probabilistic forecasts should also be explored.

6. Acknowledgments

This work could not have been performed without the support of the Navy's JTWC leadership, including especially CAPT Michael Angove and LCDR Jeremy Callahan. Special thanks go to Mr. Matthew Kucas of JTWC Techniques Development for being very generous in contributing his time, expertise and patience. Prof. Eva Regnier of the

Naval Postgraduate School contributed crucially important insights and ideas. Finally, Wei Dong generously volunteered her time and expertise to the creation and preparation of Figures 3-6, 3-7, and 3-8. Any inaccuracies or imprecision contained in this report are the sole responsibility of the author.

7. References

Bisantz, A., Kirlik, A., Gay, P., Phipps, D., Walker, N., and Fisk, A.D., (2000). Modeling and analysis of a dynamic judgment task using a lens model approach. *IEEE Transactions on Systems, Man, and Cybernetics, Vol. 30*, 6, 605-616.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217.

Cannon-Bowers, J.A. and Salas, E. (1998), *Making Decisions Under Stress: Implications for Individual and Team Training*. Washington, DC: American Psychological Association.

Carr, L.E., Elsberry, R.L. & Peak, J.E. (2001). Beta test of the systematic approach expert system prototype as a tropical cyclone track forecasting aid. *Weather and Forecasting, 16*, 355-368.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299*, 999-1001.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors, 4*. 59-73.

Elsberry, R.L. and Carr, L.E. (2000). Consensus of dynamical tropical cyclone track forecasts – Errors versus spread. *Monthly Weather Review, 128*, 4131-4138.

Goerss, J.S., Sampson, C.R. & Gross, J.M. (2004). A history of Western North Pacific tropical cyclone track forecast skill. *Weather and Forecasting, 19*, 633-638.

Horrey, W. J., Wickens, C. D., & Stewart, T. & A. Kirlik (2006). Supporting situation assessment through attention guidance and diagnostic aiding: Benefits, costs, and the impact of automation on judgment skill. In A. Kirlik (Ed.), *Adaptive Perspectives on Human-Technology Interaction* (pp. 55-70). NY: Oxford U. Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kirlik, A. (2006). *Adaptive Perspectives of Human-Technology Interaction: Methods and models for cognitive engineering and human-computer interaction*. New York: Oxford University Press.

Kirlik, A. & Strauss, R. (2006). Situation awareness as judgment I: Theoretical framework, modeling, and quantitative measurement. *International Journal of Industrial*

Ergonomics. Special Issue on New Insights in Human Performance and Decision Making, 36, 463-474.

Kirlik, A., Fisk, A.D., Walker, N. & Rothrock, L. (1998). Feedback augmentation and part-task practice in training dynamic decision making skills. In J.A. Cannon-

Bowers and E. Salas (Eds), *Making Decisions Under Stress: Implications for Individual and Team Training* (pp. 91-113). Washington, DC: American Psychological Association.

McCarley, J. & Benjamin, A. (in press). Bayesian and signal detection models. In J.D. Lee and A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering*. New York: Oxford University Press.

Miller, S., Kirlik, A., Kosorukoff, A., Tsai, J. (2008) Supporting joint human-computer judgment under uncertainty. *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors and Ergonomics Society.

Sampson, C.R., Knaff, J.A., & Fukada, E.M. (2007). Operational evaluation of a selective consensus in the Western North Pacific Basin. *Weather and Forecasting, 22*, 671-675.

Strauss, R. & Kirlik, A. (2006). Situation awareness as judgment II: Experimental evaluation and demonstration. *International Journal of Industrial Ergonomics: Special Issue on New Insights in Human Performance and Decision Making*. 36, 475-484.

Tsai, J., Miller, S. & Kirlik, A. (2011). Interactive visualizations to improve Bayesian reasoning. *Proceedings of the 2011 Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: Human Factors & Ergonomics Society.

Tsui, T.L. and Miller, R.J. (1988). Evaluation of Western North Pacific tropical cyclone objective forecast aids. *Weather Forecasting, 3*, 76-85.

Wickens, C.D., Lee, J.D., Liu, Y. & S. Becker (2004). *An Introduction to Human Factors Engineering, 2nd Edition*. Upper Saddle River, NJ: Pearson.

LIST OF REFERENCES

Brooks, H.E. (2004). Tornado-warning performance in the past and future: A perspective from signal detection theory. *Bulletin of the American Meteorological* Society, *85*(6), pp. 837-843.

DeMaria, M., Knaff, J.A., & Sampson, C. (2007). Evaluation of long-term trends in tropical cyclone intensity forecasts. *Meteorology and Atmospheric Physics*, *97*(1), pp. 19-28.

Kirlik, A. (Ed.). (2006). *Adaptive perspectives on human-technology interaction: Methods and models for cognitive engineering and human-computer interaction.* Oxford University Press, New York, NY.

Knaff, J.A., Sampson, C.R., DeMaria, M., Marchok, T., Gross, J.M., & McAdie, C.J. (2007). Statistical tropical cyclone wind radii prediction using climatology and persistence. *Weather and Forecasting*, *22*(4), pp. 781-791.

National Oceanic and Atmospheric Administration, Science Advisory Board, Hurricane Intensity Research Working Group. (2006). Majority report. Retrieved from http://www.sab.noaa.gov/Reports/HIRWG_final73.pdf

Stewart, T.R., Roebber, P.J., & Bosart, L.F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior and Human Decision Processes*, *69*(3), pp. 205-219.

Stewart, T.R. (2001). Improving reliability of judgmental forecasts. In J.S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 81-106). Kluwer Academic Publishers, New York, NY.

Stewart, T.R., & Lusk, C.M. (1994). Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. *Journal of Forecasting*, *13*(7), pp. 579-599.

Stewart, T.R. (1990) A decomposition of the correlation coefficient and its use in analyzing forecast skill. *Weather and Forecasting* *5*(4), pp. 661-666.

INITIAL DISTRIBUTION LIST

1. Research Office (Code 09)..1
Naval Postgraduate School
Monterey, CA 93943-5000

2. Dudley Knox Library (Code 013)..2
Naval Postgraduate School
Monterey, CA 93943-5002

3. Defense Technical Information Center..2
8725 John J. Kingman Rd., STE 0944
Ft. Belvoir, VA 22060-6218

4. Richard Mastowski (Technical Editor)..2
Graduate School of Operational and Information Sciences (GSOIS)
Naval Postgraduate School
Monterey, CA 93943-5219

5. CAPT Angove, USN..1
Commander, Naval Maritime Forecast Center/Joint Typhoon Warning Center
425 Luapele Road
Pearl Harbor, HI 96860

6. Alex Kirlik..1
1201 Waverly Drive
Champaign, IL 61821